No. 598 **March 2019**

Matrix-free subcell residual distribution for Bernstein finite elements: Low-order schemes and FCT

H. Hajduk, D. Kuzmin, T. Kolev, R. Abgrall

ISSN: 2190-1767

Matrix-free subcell residual distribution for Bernstein finite elements: Low-order schemes and FCT

Hennes Hajduk[a], Dmitri Kuzmin[a], Tzanio Kolev[b], Remi Abgrall[c]

[a] *Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, D-44227 Dortmund, Germany*
[b] *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, P.O. Box 808, L-561, Livermore, CA 94551, USA*
[c] *Institute of Computational Sciences, Winterthurerstrasse 190 Universität Zürich, CH-8057 Zurich, Switzerland*

Abstract

In this work, we introduce a new residual distribution (RD) framework for the design of matrix-free bound-preserving finite element schemes. As a starting point, we consider continuous and discontinuous Galerkin discretizations of the linear advection equation. To construct the corresponding local extremum diminishing (LED) approximation, we perform mass lumping and redistribute the element residuals in a manner which guarantees the LED property. The hierarchical correction procedure for high-order Bernstein finite element discretizations involves localization to subcells and definition of bound-preserving weights for subcell contributions. Using strong stability preserving (SSP) Runge-Kutta methods for time integration, we prove the validity of discrete maximum principles under CFL-like time step restrictions. The low-order version of our method has roughly the same accuracy as the one derived from a piecewise (multi)-linear approximation on a submesh with the same nodal points. In high-order extensions, we currently use a flux-corrected transport (FCT) algorithm which can also be interpreted as a nonlinear RD scheme. The properties of the algebraically corrected Galerkin discretizations are illustrated by 1D, 2D, and 3D examples for Bernstein finite elements of different order. The results are as good as those obtained with the best matrix-based approaches. In our numerical studies for multidimensional

Email addresses: hennes.hajduk@math.tu-dortmund.de (Hennes Hajduk), kuzmin@math.uni-dortmund.de (Dmitri Kuzmin), tzanio@llnl.gov (Tzanio Kolev), remi.abgrall@math.uzh.ch (Remi Abgrall)

problems, we use quadrilateral/hexahedral meshes but our methodology is readily applicable to unstructured/simplicial meshes as well.

Keywords: advection equation, discrete maximum principles, Bernstein finite elements, matrix-free methods, residual distribution, flux-corrected transport

---

1. Introduction

Numerical solution of hyperbolic equations using Galerkin methods based on high-order continuous or discontinuous finite element approximations requires implementation of certain control mechanisms for detecting and preventing violations of discrete maximum principles (DMPs). In many cases, at least local corrections of the standard Galerkin discretization are necessary, to ensure that the finite element solutions are bounded above and/or below as required by physical or numerical admissibility conditions. Depending on the design criteria and qualitative properties of the unknown exact solution, methods that are guaranteed to produce numerical solutions free of undershoots/overshoots are called monotone, positive, monotonicity-preserving, local extremum diminishing, total variation diminishing, maximum principle preserving, positivity-preserving, etc. In this article, we will generally call a numerical scheme *bound-preserving* if it satisfies local or global DMPs consistent with certain properties of the exact solution (e.g., nonnegativity or boundedness in terms of the initial/boundary values), see Section 3.

Bound-preserving finite element schemes commonly use artificial diffusion operators and/or *limiting* techniques to enforce relevant constraints. Algebraic approaches modify the matrices or residuals of the Galerkin discretization in a way which provides the desired DMP properties. Examples of such methods include *flux-corrected transport* (FCT) algorithms of predictor-corrector type [25, 29, 31], monolithic *algebraic flux correction* (AFC) schemes [10, 11, 12, 29, 33], and *residual distribution* (RD) methods [1, 2, 4, 18] (also known as *fluctuation splitting* schemes [15, 17]). In the framework of discontinuous Galerkin (DG) methods, preservation of global or local bounds is commonly enforced using some kind of flux or slope limiting [27, 28, 38]. If the piecewise-constant DG approximation is provably bound-preserving, violations of DMP constraints can be prevented by limiting the numerical fluxes or steep gradients of a high-order finite element solution.

As of this writing, the overwhelming majority of bound-preserving finite element schemes are based on low-order (at most quadratic) polynomial approximations. However, the design of arbitrary high-order extensions has been actively pursued by several research groups in recent years. In particular, the representation of finite element solutions in terms of Bernstein basis functions has led to initial high-order generalizations of both RD and FCT algorithms [5, 6, 8, 34]. As an alternative to global corrections of a high-order Galerkin discretization, localization of limiting to thin layers of cells in which violations of maximum principles (might) occur was proposed in the context of partitioned finite element spaces [30] and DG methods based on *Multi-dimensional Optimal Order Detection* (MOOD) [21].

A well-designed limiting strategy for finite elements of degree $p > 1$ should lead to a bound-preserving scheme which is at least as accurate as the $p = 1$ version for the same number of degrees of freedom (DOFs). For this reason, the use of localization procedures based on *subcell decompositions* is an essential ingredient of modern high-order extensions [7, 21, 34, 37]. Moreover, accuracy-preserving smoothness indicators are needed to avoid unnecessary limiting at smooth extrema and circumvent the second-order accuracy barrier for numerical schemes satisfying stringent DMP constraints [34, 30]. In the case of implicit time discretizations and stationary problems, the design of efficient iterative solvers for linear and nonlinear systems becomes more involved. At the same time, efficient implementation of explicit high-order finite element schemes calls for the use of approaches that avoid inversion of consistent mass matrices [3, 5] and calculation of element matrices in the process of residual assembly. The latter requirement rules out the use of artificial diffusion operators based on *discrete upwinding* [8, 34] for high-order Bernstein elements. The Rusanov scheme, which is commonly used as a building block in RD schemes [5] and FCT algorithms [25], can be implemented in a matrix-free manner but its accuracy deteriorates rapidly as the polynomial degree p is increased while keeping the total number of DOFs fixed.

The objective of this work is to develop a unified framework for matrix-free algebraic corrections of continuous Galerkin (CG) and DG methods based on high-order Bernstein finite elements. In this context, an algorithm is called matrix-free if it can be implemented without calculating global matrices or even element matrices. Such implementations are feasible if the time integration scheme is explicit or matrix-free iterative solvers are employed. If the right-hand sides and residuals of discrete problems can be calculated directly, dramatic speedups can be achieved by avoiding the overhead cost

associated with the computation of matrix entries and indirect addressing. Such matrix-free methods are also of increasing importance for modern computer architectures [14] and next-generation high-order applications [9].

As a model problem, we use the linear advection equation but the proposed methodology can be readily extended to general conservation laws. We begin with the derivation of a matrix-free nonlinear low-order scheme. Following the residual distribution approach to algebraic stabilization of Galerkin methods, we extend it to the DG framework, simplify the definition of the bound-preserving weights, propose a new subcell localization procedure, prove the desired DMP properties, and perform bound-preserving antidiffusive corrections using the element-based FCT algorithms developed in [8, 34]. In contrast to classical RD approaches based on nonlinear extensions of the Rusanov scheme, our matrix-free alternative to discrete upwinding does not require estimation of the maximum wave speed and exhibits p-independent convergence behavior with respect to the number of degrees of freedom. The accuracy of different low-order schemes and their FCT counterparts is illustrated by numerical results for 1D, 2D, and 3D test problems. In this numerical study, we consider both continuous and discontinuous finite element approximations with Bernstein polynomials of degree up to $p = 15$. Time integration is performed using explicit SSP Runge-Kutta [23] methods.

2. Galerkin discretization

Let $u = u(\mathbf{x}, t)$ be a scalar-valued function of the independent variables $\mathbf{x} = (x_1, \ldots, x_d)^T$ and $t \geq 0$. Consider the linear advection equation

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u = 0 \qquad \text{in } \Omega, \tag{1}$$

where $\mathbf{v} = \mathbf{v}(\mathbf{x})$ is a continuous velocity field and $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary $\Gamma = \partial\Omega$. The unit outward normal at a point $\mathbf{x} \in \Gamma$ is denoted by $\mathbf{n}(\mathbf{x})$. The inflow boundary of Ω is defined by

$$\Gamma_- = \{\mathbf{x} \in \Gamma \,|\, \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}.$$

The formulation of our continuous model problem is completed by imposing the initial and boundary conditions

$$u(\mathbf{x}, t) = u_{\text{in}}(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma_-, \quad t \geq 0, \tag{2}$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega. \tag{3}$$

Additionally, we make the usual assumptions which guarantee existence and uniqueness of a weak solution (see, e.g., [20]).

To discretize (1) in space, we use the standard (continuous or discontinuous) Galerkin finite element method. Let $\mathcal{T}_h$ be a computational mesh (also called *triangulation*) composed of $E_h = |\mathcal{T}_h|$ *elements*. For simplicity, we approximate u using polynomials of the same degree $p \geq 1$ in each element. Let $\varphi_1^e, \ldots, \varphi_N^e$ be the nonnegative Bernstein basis functions spanning the corresponding polynomial space on $K^e \in \mathcal{T}_h$. The definition of these basis functions for simplicial and tensor product meshes, as well as an in-depth description of their properties, can be found, e.g., in [34]. The restriction of the numerical solution u_h to K^e is given by

$$u_h^e(\mathbf{x}, t) = \sum_{j=1}^{N} u_j^e(t)\varphi_j^e(\mathbf{x}), \qquad \mathbf{x} \in K^e, \quad t \geq 0. \tag{4}$$

The local degree of freedom u_j^e is the coefficient multiplying the Bernstein basis function $\varphi_j^e : K^e \to [0, 1]$ which attains its maximum at the j-th nodal point $\mathbf{x}_j^e$ of K^e. Its global number $i_j^e = \mathcal{I}^e(j)$ can be retrieved using the DOF mapping

$$\mathcal{I}^e : \{1, \ldots, N\} \quad \to \quad \{i_1^e, \ldots, i_N^e\} =: \mathcal{N}^e, \qquad j \quad \mapsto \quad i_j^e.$$

The set of elements containing a node with the global number $i \in \{1, \ldots, N_h\}$ is denoted by $\mathcal{E}_i$. In the discontinuous Galerkin version, the Bernstein coefficients corresponding to the internal and external traces of u_h on $\partial K^e \backslash \Gamma$ receive different global node numbers. Hence, in the DG case the set $\mathcal{E}_i$ consists of a single element and $\mathcal{N}^e \cap \mathcal{N}^{e'} = \emptyset$ for any pair of different elements $K^e, K^{e'} \in \mathcal{T}_h$. In the case of a globally continuous approximation u_h, different local mappings $\mathcal{I}^e$ may produce the same global index i.

The global index notation is used, e.g., for the piecewise-polynomial global basis functions $\varphi_i,\ i = 1, \ldots, N_h$, which satisfy $\varphi_i|_{K^e} = \varphi_j^e$ when $\mathcal{I}^e(j) = i$. If global node numbers are used in the same equation as local basis functions φ_j^e or linear combinations thereof, the required index conversion is performed automatically. For example, if φ_i appears in the same formula as u_h^e defined by (4), then the subscript i is just the shorthand notation for $\mathcal{I}^e(j)$. Using this convention, the Galerkin discretization of problem (1) can be written as

$$\sum_{K^e \in \mathcal{T}_h} \int_{K^e} \varphi_i \left(\frac{\partial u_h^e}{\partial t} + \mathbf{v} \cdot \nabla u_h^e \right) \mathrm{d}\mathbf{x} + \sum_{K^e \in \mathcal{T}_h} \int_{\partial K^e} \varphi_i (\hat{u}_h^e - u_h^e) \mathbf{v} \cdot \mathbf{n}^e \mathrm{d}s = 0, \tag{5}$$

where $\mathbf{n}^e$ is the unit outward normal to ∂K^e and

$$\hat{u}_h^e(\mathbf{x},t) = \begin{cases} \lim\limits_{\epsilon\to+0} u_h(\mathbf{x}+\epsilon\mathbf{n}^e(\mathbf{x})) & \text{if } \mathbf{x}\in\partial K^e\backslash\Gamma_-,\ \mathbf{v}(\mathbf{x})\cdot\mathbf{n}^e(\mathbf{x})<0, \\ \lim\limits_{\epsilon\to+0} u_h(\mathbf{x}-\epsilon\mathbf{n}^e(\mathbf{x})) & \text{if } \mathbf{x}\in\partial K^e,\ \mathbf{v}(\mathbf{x})\cdot\mathbf{n}^e(\mathbf{x})\geq 0, \\ u_{\text{in}}(\mathbf{x},t) & \text{if } \mathbf{x}\in\partial K^e\cap\Gamma_-. \end{cases} \tag{6}$$

If u_h is globally continuous in $\bar{\Omega}$, then $\hat{u}_h^e = u_h^e$ on $\partial K^e\backslash\Gamma_-$ and

$$\sum_{K^e\in\mathcal{T}_h}\int_{\partial K^e}\varphi_i(\hat{u}_h^e-u_h^e)\mathbf{v}\cdot\mathbf{n}^e\mathrm{d}s = \int_{\Gamma_-}\varphi_i(u_{\text{in}}-u_h)\mathbf{v}\cdot\mathbf{n}\mathrm{d}s. \tag{7}$$

In DG methods, $\hat{u}_h^e$ is the upwind-sided trace which equals u_h^e for $\mathbf{x}\in\partial K^e$ with $\mathbf{v}(\mathbf{x})\cdot\mathbf{n}^e(\mathbf{x})\geq 0$. Hence, the integral over ∂K^e reduces to

$$\int_{\partial K^e}\varphi_i(\hat{u}_h^e-u_h^e)\mathbf{v}\cdot\mathbf{n}^e\mathrm{d}s = \int_{\partial K^e}\varphi_i(\hat{u}_h^e-u_h^e)\min\{0,\mathbf{v}\cdot\mathbf{n}^e\}\mathrm{d}s \tag{8}$$

and can be interpreted as a penalty term incorporating a weakly imposed inflow boundary condition for the local problem associated with K^e.

The set of edges or faces F that form the boundary ∂K^e of element K^e will be denoted by $\mathcal{F}^e$. Let $\mathcal{J}^e\subset\{1,\ldots,N\}$ be the set of nodes belonging to $\partial K^e = \bigcup_{F\in\mathcal{F}^e} F$. The internal trace $u_h^e|_F$ and its external counterpart $\hat{u}_h^e|_F$ can be written in terms of the same basis functions $\{\varphi_i^e \,|\, i\in\mathcal{J}^e\}$. The corresponding Bernstein coefficients will be denoted by u_i^e and $\hat{u}_i^e$, respectively. In vectors $\hat{u}^e = \{\hat{u}_i^e\}_{i=1}^N$ that we sometimes use in representations of boundary terms as matrix-vector products, we set $\hat{u}_i^e = u_i^e$ for $i\in\{1,\ldots,N\}\backslash\mathcal{J}^e$.

Decomposing integrals over ∂K^e into sums of integrals over $F\in\mathcal{F}^e$ and summing over $K^e\in\mathcal{T}_h$, we arrive at the DG generalization of (7),

$$\sum_{K^e\in\mathcal{T}_h}\int_{\partial K^e}\varphi_i(\hat{u}_h^e-u_h^e)\mathbf{v}\cdot\mathbf{n}^e\mathrm{d}s = \int_{\Gamma_-}\varphi_i(u_{\text{in}}-u_h)\mathbf{v}\cdot\mathbf{n}\mathrm{d}s + \sum_{F\in\mathcal{F}_h}\int_F\varphi_i[\![u_h]\!]\mathbf{v}\cdot\mathbf{n}\mathrm{d}s,$$

where $\mathcal{F}_h$ is the set of internal boundaries and $[\![u_h]\!]\mathbf{v}\cdot\mathbf{n}$ is the jump of the convective flux across the edge/face $F\in\mathcal{F}^e$.

Substitution of (4) into (5) yields the system of semi-discrete equations

$$\sum_{j=1}^{N_h} m_{ij}\frac{\mathrm{d}u_j}{\mathrm{d}t} = \rho_i^H + \sigma_i^H, \qquad i=1,\ldots,N_h \tag{9}$$

for the time-dependent Bernstein coefficients $u_i(t)$ of the high-order (superscript H) Galerkin approximation. By definition, we have

$$m_{ij} = \sum_{e \in \mathcal{E}_i \cap \mathcal{E}_j} m_{ij}^e, \quad m_{ij}^e = \int_{K^e} \varphi_i^e \varphi_j^e \mathrm{d}\mathbf{x}, \quad \sum_{j=1}^N m_{ij}^e u_j^e = \int_{K^e} \varphi_i^e u_h^e \mathrm{d}\mathbf{x}, \tag{10}$$

where m_{ij} is an entry of the consistent mass matrix $M_C = \{m_{ij}\}_{i,j=1}^{N_h}$. The contribution of K^e to the right-hand side of (9) is given by

$$\rho_i^H = \sum_{e \in \mathcal{E}_i} \rho_i^{e,H}, \qquad \sigma_i^H = \sum_{e \in \mathcal{E}_i} \sigma_i^{e,H}, \tag{11}$$

$$\rho_i^{e,H} = -\int_{K^e} \varphi_i^e \mathbf{v} \cdot \nabla u_h^e \mathrm{d}\mathbf{x} = \sum_{j=1}^N c_{ij}^e u_j^e, \qquad c_{ij}^e = -\int_{K^e} \varphi_i^e \mathbf{v} \cdot \nabla \varphi_j^e \mathrm{d}\mathbf{x}, \tag{12}$$

$$\sigma_i^{e,H} = -\sum_{F \in \mathcal{F}^e} \int_F \varphi_i^e (\hat{u}_h^e - u_h^e) \min\{0, \mathbf{v} \cdot \mathbf{n}^e\} \mathrm{d}s = \sum_{j=1}^N s_{ij}^e (\hat{u}_j^e - u_j^e), \tag{13}$$

$$s_{ij}^e = -\sum_{F \in \mathcal{F}^e} \int_F \varphi_i^e \varphi_j^e \min\{0, \mathbf{v} \cdot \mathbf{n}^e\} \mathrm{d}s. \tag{14}$$

In the CG version, the integrals over $F \in \mathcal{F}_h$ vanish because the one-sided limits $\hat{u}_h^e$ and u_h^e of the finite element solution coincide on $\partial K^e \backslash \Gamma_-$.

For conciseness, we will sometimes consider element matrices and vectors containing all contributions of K^e to system (9). For example, the consistent element mass matrix is defined by $M_C^e = \{m_{ij}^e\}_{i,j=1}^N$. The element contribution to the left-hand side of (9) is given by $M_C^e \frac{\mathrm{d}}{\mathrm{d}t} u^e$, where $u^e = \{u_i^e\}_{i=1}^N$ is the vector of local degrees of freedom. The element vectors $\rho^{e,H} = \{\rho_i^{e,H}\}_{i=1}^N$ and $\sigma^{e,H} = \{\sigma_i^{e,H}\}_{i=1}^N$ can be written as $\rho^{e,H} = C^e u^e$ and $\sigma^{e,H} = S^e(\hat{u}^e - u^e)$, where $C^e = \{c_{ij}^e\}_{i,j=1}^N$ and $S^e = \{s_{ij}^e\}_{i,j=1}^N$ are element matrices. However, we emphasize that the knowledge of element matrices is **not** required for practical implementation purposes because element vectors representing contributions of K^e to the residual of (9) do not need to be expressed as matrix-vector products and can be calculated efficiently in a matrix-free manner. In matrix-based implementations, the sparsity of S^e should be taken into account. Note that $s_{ij}^e = 0$ unless nodes i and j belong to the same edge/face $F \in \mathcal{F}^e$ and $\mathbf{v} \cdot \mathbf{n}^e < 0$ on a non-empty subset of F.

3. Bound-preserving schemes

Since the Bernstein basis functions φ_j^e are nonnegative and form a partition of unity [34], the Bernstein polynomial u_h^e defined by (5) is bounded in terms of the local degrees of freedom as follows:

$$u_{\min}^e := \min_{1\le j\le N} u_j^e \le u_h^e(\mathbf{x}) \le \max_{1\le j\le N} u_j^e =: u_{\max}^e \qquad \forall \mathbf{x} \in K^e. \tag{15}$$

Therefore, the range of solution values is determined by the maxima and minima of the Bernstein coefficients. This is an important advantage of the Bernstein basis (e.g. compared to nodal bases) which makes it possible to constrain u_h in a desired manner by constraining its DOFs. For example, a finite element scheme which is positivity-preserving with respect to the Bernstein DOFs cannot produce negative solution values anywhere.

It is well known that the standard Galerkin discretization of the linear advection equation is not bound-preserving. Therefore, we will modify it using the design criteria presented in this section.

Consider a numerical scheme that leads to the semi-discrete problem

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = f_i, \qquad i = 1, \dots, N_h, \tag{16}$$

where f_i is a Lipschitz-continuous function of the DOFs u_j, $j \in \mathcal{N}^e$, $e \in \mathcal{E}_i$ and $m_i = \sum_{j=1}^{N_h} m_{ij} = \int_\Omega \varphi_i \mathrm{d}\mathbf{x}$ is a diagonal entry of the lumped mass matrix $M_L = \{\delta_{ij} m_i\}_{i,j=1}^{N_h}$. Due to the nonnegativity of φ_i, we have $m_i > 0$. This is another important advantage of the Bernstein basis representation compared to high-order Lagrange finite element approximations, for which row-sum lumping can produce zero diagonal entries.

If time integration is performed using an explicit SSP Runge-Kutta method [23], each stage corresponds to a forward Euler update of the form

$$m_i \tilde{u}_i = m_i u_i + \Delta t f_i(u), \qquad i = 1, \dots, N_h, \tag{17}$$

where $u = \{u_i\}_{i=1}^{N_h}$ is the vector of bound-preserving DOFs calculated at the previous time step or Runge-Kutta stage and Δt is the time increment.

We say that such an update is *local extremum diminishing (LED)* if

$$u_{i,\min} \le \tilde{u}_i \le u_{i,\max} \qquad \forall i = 1, \dots, N_h, \tag{18}$$

where $u_{i,\min}$ and $u_{i,\max}$ are defined in terms of $u^e_{\min}$ and $u^e_{\max}$ as follows:

$$u_{i,\min} = \min_{e\in\bar{\mathcal{E}}_i} u^e_{\min}, \qquad u_{i,\max} = \max_{e\in\bar{\mathcal{E}}_i} u^e_{\max}. \tag{19}$$

The set $\bar{\mathcal{E}}_i$ contains the numbers of all elements to which the point $\mathbf{x}_i$ belongs. If $\mathbf{x}_i$ is an internal node of $K^e \in \mathcal{T}_h$, then $\bar{\mathcal{E}}_i = \{e\}$. If $\mathbf{x}_i$ is an internal node of $F = \partial K^e \cap \partial K^{e'}$ for $K^e, K^{e'} \in \mathcal{T}_h$, $K^e \neq K^{e'}$, then $\bar{\mathcal{E}}_i = \{e, e'\}$. If $\mathbf{x}_i$ is a node on the boundary of F (i.e., a point on the edge of a 3D element or a vertex), then the numbers of all elements meeting at this point are included in the set $\bar{\mathcal{E}}_i$ of host elements. This definition of $\bar{\mathcal{E}}_i$ in (19) implies that the same bounds are used for all Bernstein coefficients u^e_i associated with the same space location $\mathbf{x}_i$. In the CG version, we have $\bar{\mathcal{E}}_i = \mathcal{E}_i$.

Theorem 1 (LED criterion). *An update of the form (17) is LED if*

$$\kappa_i(u_{i,\min} - u_i) \le f_i(u) \le \kappa_i(u_{i,\max} - u_i) \tag{20}$$

for some bounded coefficients $\kappa_i \ge 0$ and the time step Δt satisfies

$$\frac{\kappa_i \Delta t}{m_i} \le 1 \qquad \forall i = 1, \dots, N_h. \tag{21}$$

Proof. If condition (20) holds, then $f_i(u) = \beta_i \kappa_i (u_k - u_i)$, where

$$u_k = \begin{cases} u_{i,\max} & \text{if } f_i(u) > 0, \\ u_{i,\min} & \text{if } f_i(u) \le 0, \end{cases}$$

and $\beta_i \in [0, 1]$. Condition (21) implies $0 \le \nu_i \le 1$ for the "CFL number" $\nu_i = \frac{\Delta t}{m_i}\beta_i\kappa_i$. It follows that the new value $\tilde{u}_i = u_i + \frac{\Delta t}{m_i} f_i(u) = u_i + \nu_i(u_k - u_i)$ $= (1 - \nu_i)u_i + \nu_i u_k$ is a convex combination of u_i and u_k, both of which are bounded by the local extrema $u_{i,\min}$ and $u_{i,\max}$. □

Theorem 1 implies that scheme (17) is LED for $\Delta t > 0$ if κ_i defined by

$$\kappa_i = \begin{cases} \frac{f_i}{u_{i,\max} - u_i} & \text{if } f_i > 0, \\ \frac{f_i}{u_{i,\min} - u_i} & \text{if } f_i < 0, \\ 0 & \text{if } f_i = 0 \end{cases} \tag{22}$$

is bounded. For practical design of LED finite element schemes, we formulate localized sufficient conditions which provide this desirable property.

Theorem 2 (Localized LED criterion). *Consider system (16) with*

$$f_i(u) = \sum_{e \in \mathcal{E}_i} f_i^e, \qquad i = 1, \dots, N_h, \tag{23}$$

where f_i^e are element contributions such that the coefficients

$$\kappa_i^e = \begin{cases} \frac{f_i^e}{u_{i,\max} - u_i^e} & \text{if } f_i^e > 0, \\ \frac{f_i^e}{u_{i,\min} - u_i^e} & \text{if } f_i^e < 0, \\ 0 & \text{if } f_i^e = 0 \end{cases} \tag{24}$$

are bounded for all $e \in \mathcal{E}_i$. Then update (17) is LED for time steps Δt satisfying the CFL-like condition (21) with

$$\kappa_i = \sum_{e \in \mathcal{E}_i} \kappa_i^e. \tag{25}$$

Proof. We have $f_i^e = \kappa_i^e(u_{i,\max}^e - u_i^e)$ or $f_i^e = \kappa_i^e(u_{i,\min}^e - u_i^e)$ by (24). Hence, $f_i(u)$ defined by (23) satisfies (20) with κ_i given by (25). □

Below we show that the LED property holds (under time step restrictions) for $f_i(u)$ assembled from element contributions of the form $f^e = \rho^e + \sigma^e$, where

$$\rho^e = \tilde{C}^e u^e, \qquad \sigma^e = \tilde{S}^e(\hat{u}^e - u^e), \tag{26}$$

and the element matrices $\tilde{C}^e \in \mathbb{R}^{N \times N}$ and $\tilde{S}^e \in \mathbb{R}^{N \times N}$ satisfy

$$\sum_{j=1}^{N} \tilde{c}_{ij}^e = 0, \qquad \tilde{c}_{ij}^e \geq 0 \quad \forall j \in \{1, \dots, N\} \backslash \{i\}, \tag{27}$$

$$\tilde{s}_{ij}^e = \delta_{ij} s_i^e, \qquad s_i^e \geq 0. \tag{28}$$

Indeed, individual components f_i^e of the element vector f^e can be written as

$$f_i^e = \sum_{\substack{j=1 \\ j \neq i}}^{N} \tilde{c}_{ij}^e (u_j^e - u_i^e) + s_i^e(\hat{u}_i^e - u_i^e) = \kappa_i^e(u_k - u_i^e),$$

where $\kappa_i^e \geq 0$ is bounded and $u_k \in \{u_{i,\min}, u_{i,\max}\}$ is a local extremum.

Let $m_i^e = \sum_{j=1}^N m_{ij}^e > 0$ be the i-th diagonal entry of the lumped element mass matrix $M_L^e = \{\delta_{ij} m_i^e\}_{i,j=1}^N$. Introducing $\tilde{u}_i^e = u_i^e + \frac{\Delta t}{m_i^e} f_i^e$ and following the proof of Theorem 1, we find that $u_{i,\min} \le \tilde{u}_i^e \le u_{i,\max}$ for $\Delta t \kappa_i^e \le m_i^e$. In the DG version, $\tilde{u}_i = \tilde{u}_i^e$ is the final result since the set $\mathcal{E}_i = \{e\}$ contains exactly one element number and node i receives just the element contribution f_i^e. In the CG version, node i may belong to multiple elements and the uniquely defined Bernstein coefficient $\tilde{u}_i$ satisfies (cf. [16, 34])

$$m_i \tilde{u}_i = \sum_{e \in \mathcal{E}_i} m_i^e \tilde{u}_i = \sum_{e \in \mathcal{E}_i} (m_i^e u_i^e + \Delta t f_i^e) = \sum_{e \in \mathcal{E}_i} m_i^e \tilde{u}_i^e.$$

Hence, $\tilde{u}_i$ is a convex combination of the one-sided approximations $\tilde{u}_i^e$ which possesses the LED property under the same time step restrictions.

Theorem 3 (LED corrections of element contributions). *Consider*

$$f^e = \rho^e + \sigma^e + \eta^e, \tag{29}$$

where ρ^e and σ^e satisfy the conditions of Theorem 2. Then the LED property is preserved under the time step restriction of Theorem 2 if

$$\frac{m_i^e}{\Delta t}(u_{i,\min} - \tilde{u}_i^e) \le \eta_i^e \le \frac{m_i^e}{\Delta t}(u_{i,\max} - \tilde{u}_i^e), \tag{30}$$

where $\tilde{u}_i^e$ is the bound-preserving one-sided approximation defined by

$$M_L^e \tilde{u}^e = M_L^e u^e + \Delta t(\rho^e + \sigma^e).$$

Proof. The corrected values $\bar{u}_i$ of the Bernstein coefficients are defined by

$$\sum_{e \in \mathcal{E}_i} m_i^e \bar{u}_i = \sum_{e \in \mathcal{E}_i} (m_i^e u_i^e + \Delta t f_i^e) = \sum_{e \in \mathcal{E}_i} (m_i^e \tilde{u}_i^e + \Delta t \eta_i^e) = \sum_{e \in \mathcal{E}_i} m_i^e \bar{u}_i^e,$$

where $\bar{u}_i^e = \tilde{u}_i^e + \frac{\Delta t}{m_i^e} \eta_i^e$ stay in the range $[u_{i,\min}, u_{i,\max}]$ if condition (30) holds. □

Theorem 3 provides a convenient tool for the design of flux-corrected transport (FCT) algorithms [8, 25, 31, 34]. Following the FCT approach, we will (i) modify the element contributions of the high-order Galerkin method so as to satisfy the conditions of Theorem 2 and (ii) correct the resulting low-order scheme by adding η^e limited as required by Theorem 3.

4. Linear low-order schemes

Replacing the consistent mass matrix M_C by its lumped counterpart M_L, we transform (9) into a system of ODEs of the form (16). Since the Galerkin element contributions $f^{e,H} = C^e u^e + S^e(\hat{u}^e - u^e)$ are not of LED type, the element vectors $\rho^{e,H} = C^e u^e$ and $\sigma^{e,H} = S^e(\hat{u}^e - u^e)$ need to be replaced by their LED counterparts $\rho^{e,L}$ and $\sigma^{e,L}$. The superscript L indicates that the resulting approximations will be of low order in accordance with the Godunov theorem [22]. The modified scheme will remain conservative if

$$\sum_{i=1}^{N} \rho_i^{e,H} = \sum_{i=1}^{N} \rho_i^{e,L}, \qquad \sum_{i=1}^{N} \sigma_i^{e,H} = \sum_{i=1}^{N} \sigma_i^{e,L}. \tag{31}$$

In the 1D case, the element matrix $\tilde{S}^e := S^e$ satisfies (28). Hence, the boundary terms $\sigma_i^{e,L} = \sigma_i^{e,H}$ do not endanger the LED property. To enforce it in 2D or 3D, the diagonal matrix $\tilde{S}^e = \{\delta_{ij} s_i^e\}_{i,j=1}^N$ with

$$s_i^e = \sum_{j=1}^{N} s_{ij}^e = -\sum_{F \in \mathcal{F}^e} \int_F \varphi_i^e \min\{0, \mathbf{v} \cdot \mathbf{n}^e\} \mathrm{ds} \tag{32}$$

can readily be constructed using "mass lumping" for boundary terms. The corresponding low-order element contributions are given by

$$\sigma_i^{e,L} = s_i^e(\hat{u}_i^e - u_i^e). \tag{33}$$

The conservation property $\sum_{i=1}^N \sigma_i^{e,H} = \sum_{i=1}^N s_i^e(\hat{u}_i^e - u_i^e) = \sum_{i=1}^N \sigma_i^{e,L}$ follows from the fact that $\sum_{i=1}^N \varphi_i^e \equiv 1$ and $s_{ij}^e = s_{ji}^e$ for $i, j = 1, \ldots, N$.

The element matrix $\tilde{C}^e$ of a low-order LED scheme can be defined by adding a discrete diffusion operator $D^e = \{d_{ij}^e\}_{i,j=1}^N$ to C^e. The discrete conservation property is preserved if D^e has zero column sums. The LED condition (27) holds if $\tilde{C}^e = C^e + D^e$ has zero row sums and $\tilde{c}_{ij}^e \geq 0$ for all $j \neq i$. For low-order finite elements (linear or multilinear, $p = 1$), the least diffusive LED approximation of this kind is defined by [12, 29, 31]

$$d_{ij}^e = \begin{cases} \max\{-c_{ij}^e, 0, -c_{ji}^e\} & \text{if } j \neq i, \\ -\sum\limits_{\substack{k=1 \\ k \neq i}}^{N} d_{ik}^e & \text{if } j = i. \end{cases} \tag{34}$$

As shown in [12, 33] for continuous finite elements, its provable order of accuracy is $\frac{1}{2}$ on general meshes. This approach to the definition of

$$\rho^{e,L} = \tilde{C}^e u^e, \qquad \tilde{C}^e = C^e + D^e \tag{35}$$

is known as *discrete upwinding* [29]. It is readily applicable to Bernstein elements of any degree but low-order solutions obtained using formula (34) become less accurate as the degree p of Bernstein polynomials is increased while keeping the total number of DOFs fixed. As shown in [34], this side effect can be avoided (for constant velocity fields) or significantly alleviated (for general velocity fields) by using $\tilde{C}^e = P^e C^e + \tilde{D}^e$, where $P^e = M_L^e (M_C^e)^{-1}$ is the local mass lumping operator and $\tilde{D}^e$ is the artificial operator constructed using discrete upwinding for $P^e C^e$. This "preconditioned" version yields

$$\rho^{e,L} = \rho^{e,H} + (M_L^e - M_C^e)(M_C^e)^{-1} C^e u^e + \tilde{D}^e u^e \tag{36}$$

and is conservative since components of matrix-vector products involving discrete diffusion operators (i.e., symmetric matrices with zero row and column sums, as defined in [29, 31, 33]) like $M_L^e - M_C^e$ and $\tilde{D}^e$ sum to zero.

In contrast to the Galerkin residuals $\rho^{e,H}$, the element contributions $\rho^{e,L}$ of discrete upwinding schemes cannot be calculated in a matrix-free manner. Indeed, the element matrix C^e is required for calculation of the artificial diffusion coefficients d_{ij}^e using formula (34), whereas M_C^e is additionally required to compute $\rho^{e,L}$ defined by (36).

The need to calculate an optimal diffusion coefficient d_{ij}^e for each pair of local DOFs can be avoided by using

$$d_{ij}^e = \begin{cases} d^e & \text{if } j \neq i, \\ -(N-1)d^e & \text{if } j = i, \end{cases} \tag{37}$$

where d^e is an element-based diffusion coefficient such that $\tilde{c}_{ij}^e = c_{ij}^e + d^e \geq 0$ for all $j \neq i$. Discrete diffusion operators of this kind were employed, e.g., in [25]. The corresponding low-order element contributions can be written as

$$\rho^{e,L} = \rho^{e,H} + d^e N(\bar{u}^e - u^e), \tag{38}$$

where $\bar{u}^e = \frac{1}{N}\sum_{j=1}^N u_j^e$ is the arithmetic mean of the local DOFs. Obviously, this modification of $\rho^{e,H}$ is well suited for a matrix-free implementation.

Using the Cauchy-Schwarz inequality to estimate the magnitude of c_{ij}^e, we find that the LED property of $\tilde{c}_{ij}^e$ can be enforced using

$$d^e = \max_{1\le i\le N} \|\varphi_i^e\|_{L^2(K^e)} \max_{1\le j\le N} \|\mathbf{v}\cdot\nabla\varphi_j^e\|_{L^2(K^e)}. \tag{39}$$

In the literature on residual distribution (RD) methods, the matrix-free low-order method defined by (38) is known as the *Rusanov scheme* [5, 7]. It is simple and efficient but its accuracy deteriorates rapidly as the polynomial degree p increases. Indeed, using the same artificial diffusion d^e for all off-diagonal entries (even those with very small magnitudes) produces significant amounts of numerical diffusion as the size of the element matrix increases and estimates of $|c_{ij}^e|$ become very pessimistic. For that reason, the Rusanov scheme is not to be recommended as such. However, it can be used to define the weights for more accurate nonlinear RD schemes as we demonstrate below.

5. Nonlinear low-order schemes

The linear LED schemes presented so far are either matrix-based or too inaccurate (especially for high-order Bernstein FEM). Using the framework of residual distribution methods [1, 18], we will derive nonlinear matrix-free LED schemes which exhibit p-independent convergence behavior. In this paper, we call a LED finite element approximation an RD scheme if it replaces the Galerkin element contributions $\rho_i^{e,H}$ with element contributions $\rho_i^{e,L}$ having the same total *fluctuation* $\rho_*^e = \sum_{i=1}^N \rho_i^{e,H} = \sum_{i=1}^N \rho_i^{e,L}$. Adopting this design principle, we generalize and simplify the nonlinear positive streamwise invariant (PSI) correction [1, 13, 36] of the Rusanov scheme. To achieve the desired p-invariant convergence behavior, we localize the process of residual distribution to subcells in the next section. High-order extensions of the resulting schemes are presented in Section 7.

In our new approach to residual distribution, we decompose ρ_*^e into

$$\rho_{*,+}^e = \sum_{i=1}^N \max\{0, \rho_i^{e,H}\}, \qquad \rho_{*,-}^e = \sum_{i=1}^N \min\{0, \rho_i^{e,H}\} \tag{40}$$

and require preservation of $\rho_{*,\pm}^e$. This is generally a more stringent requirement than preservation of ρ_*^e. For linear elements and constant velocity fields, we have $\rho_i^{e,H} = \frac{\rho_*^e}{N}$. Therefore, either $\rho_{*,+}^e = 0$ or $\rho_{*,-}^e = 0$ in this case.

Let $\Phi_{i,\pm}^e$ be generic distribution weights (to be defined later) satisfying

$$\tilde{\kappa}_i^e(u_{\min}^e - u_i^e) \le \Phi_{i,-}^e \le 0 \le \Phi_{i,+}^e \le \tilde{\kappa}_i^e(u_{\max}^e - u_i^e) \tag{41}$$

for some bounded $\tilde{\kappa}_i^e \ge 0$. We additionally require that

$$\sum_{j=1}^N \Phi_{j,\pm}^e = 0 \quad \Rightarrow \quad \rho_{*,\pm}^e = 0. \tag{42}$$

Given a set of weights $\Phi_{i,\pm}^e$ satisfying these conditions, we use them to redistribute the fluctuations $\rho_{*,\pm}^e$ among the nodes of K^e. Specifically, the LED element contributions of our element-based low-order RD scheme are defined by

$$\rho_i^{e,L} = \frac{\rho_{*,+}^e \Phi_{i,+}^e}{\sum_{j=1}^N \Phi_{j,+}^e + \epsilon} + \frac{\rho_{*,-}^e \Phi_{i,-}^e}{\sum_{j=1}^N \Phi_{j,-}^e - \epsilon}, \tag{43}$$

where $\epsilon > 0$ is a small constant that we use to prevent division by zero.

Remark 1. To avoid introducing the parameter ϵ, the compact representation (43) of $\rho_i^{e,L}$ can be replaced with the more formal definition

$$\rho_i^{e,L} = \begin{cases} \frac{\rho_{*,+}^e \Phi_{i,+}^e}{\sum_{j=1}^N \Phi_{j,+}^e} & \text{if } \sum_{j=1}^N \Phi_{j,+}^e > 0 \ \wedge \sum_{j=1}^N \Phi_{j,-}^e = 0, \\ \frac{\rho_{*,+}^e \Phi_{i,+}^e}{\sum_{j=1}^N \Phi_{j,+}^e} + \frac{\rho_{*,-}^e \Phi_{i,-}^e}{\sum_{j=1}^N \Phi_{j,-}^e} & \text{if } \sum_{j=1}^N \Phi_{j,+}^e > 0 \ \wedge \sum_{j=1}^N \Phi_{j,-}^e < 0, \\ \frac{\rho_{*,-}^e \Phi_{i,-}^e}{\sum_{j=1}^N \Phi_{j,-}^e} & \text{if } \sum_{j=1}^N \Phi_{j,+}^e = 0 \ \wedge \sum_{j=1}^N \Phi_{j,-}^e < 0, \\ 0 & \text{if } \sum_{j=1}^N \Phi_{j,+}^e = 0 = \sum_{j=1}^N \Phi_{j,-}^e. \end{cases}$$

Mathematically speaking, the latter definition is equivalent to (43) in the limit $\epsilon \searrow 0$. In practice, implementations based on the if-version are preferable because the use of regularization constants in the denominators leads to a lack of exact mass conservation. We adopt the ϵ notation in this paper just to avoid indeterminancies of the form $\frac{0}{0}$ without distinguishing between the cases of vanishing and nonvanishing denominators.

The element contributions $\rho_i^{e,L}$ of the generic RD scheme defined by (43) satisfy the localized LED criterion (Theorem 2) if the coefficients

$$\kappa_+^e = \frac{\rho_{*,+}^e}{\sum_{j=1}^N \Phi_{j,+}^e + \epsilon} \ge 0, \qquad \kappa_-^e = \frac{\rho_{*,-}^e}{\sum_{j=1}^N \Phi_{j,-}^e - \epsilon} \ge 0 \tag{44}$$

of the equivalent representation $\rho_i^{e,L} = \kappa_+^e \Phi_{i,+}^e + \kappa_-^e \Phi_{i,-}^e$ stay bounded for $\epsilon \searrow 0$. That is why the weights $\Phi_{i,\pm}^e$ should be chosen to satisfy (41).

Interestingly enough, the RD schemes of Abgrall et al. [5, 7] can be written in the form (43) with fluctuations $\rho_{*,+}^e = \max\{0, \rho_*^e\}$, $\rho_{*,-}^e = \min\{0, \rho_*^e\}$ and weights $\Phi_{i,+}^e = \max\{0, \tilde{\rho}_i^{e,L}\}$, $\Phi_{i,-}^e = \min\{0, \tilde{\rho}_i^{e,L}\}$, where $\tilde{\rho}_i^{e,L}$ are the LED element contributions of the linear Rusanov scheme. The replacement of $\tilde{\rho}_i^{e,L}$ with the element contributions $\rho_i^{e,L}$ defined by (43) corresponds to the PSI correction, a detailed analysis of which can be found in [13].

Remark 2. The PSI-corrected Rusanov scheme can be readily extended to general conservation laws and has proved its worth in applications to gas dynamics. We remark, however, that its original implementation in the context of SSP Runge-Kutta methods does not express the final solution as a convex combination of bound-preserving forward Euler updates. Instead, the nonlinear PSI correction is applied to time-discretized residuals that depend on more than one intermediate solution (see [5] for a detailed description of the second-order method). The resulting nonlinear schemes are essentially nonoscillatory and more accurate than SSP implementations in which residual distribution is performed for element contributions to the right-hand side of system (16) **before** invoking the Runge-Kutta time integrator. We favor the latter approach because it guarantees the LED property, whereas high accuracy can be achieved using FCT algorithms (see Section 7).

To avoid the need for calculating the Rusanov element contributions $\tilde{\rho}_i^{e,L}$ and estimating the magnitude of c_{ij}^e to obtain a lower bound for the artificial diffusion coefficient d^e, we adopt the simplest choice

$$\Phi_{i,+}^e := u_{\max}^e - u_i^e, \qquad \Phi_{i,-}^e := u_{\min}^e - u_i^e \tag{45}$$

which yields

$$\rho_i^{e,L} = \kappa_+^e (u_{\max}^e - u_i^e) + \kappa_-^e (u_{\min}^e - u_i^e), \tag{46}$$

$$\kappa_+^e = \frac{\rho_{*,+}^e}{\sum_{j=1}^N (u_{\max}^e - u_j^e) + \epsilon} \geq 0, \qquad \kappa_-^e = \frac{\rho_{*,-}^e}{\sum_{j=1}^N (u_{\min}^e - u_j^e) - \epsilon} \geq 0. \tag{47}$$

The so-defined nonlinear RD scheme can be interpreted as a simplification of the one based on the Rusanov residuals. It provides the discrete conservation property since $\sum_{i=1}^N \rho_i^{e,H} = \rho_*^e = \sum_{i=1}^N \rho_i^{e,L}$ by construction.

In the next theorem, we prove that $\kappa_{\pm}^e$ are bounded. Let the boundary terms $\sigma_i^{e,L}$ be defined by (33). Recalling definition (19) of the nodal bounds in terms of $u_{\max}^e$ and $u_{\min}^e$, and using Theorem 2 with

$$\kappa_i^e = \begin{cases} \kappa_+^e + s_i^e & \text{if } \rho_i^{e,L} > 0, \\ \kappa_-^e + s_i^e & \text{if } \rho_i^{e,L} < 0, \\ s_i^e & \text{if } \rho_i^{e,L} = 0, \end{cases} \tag{48}$$

we find that the nonlinear low-order RD scheme with $f_i^e = \rho_i^{e,L} + \sigma_i^{e,L}$ is LED for time steps satisfying (21) with $\kappa_i = \sum_{e \in \mathcal{E}_i} \kappa_i^e$. The computable sharp CFL upper bound for Δt corresponds to κ_i defined by (22).

Theorem 4. *The LED distribution coefficients $\kappa_{\pm}^e$ defined by (47) are bounded by a constant $\kappa = \kappa(e, N) > 0$.*

Proof. Since the Bernstein basis functions φ_i^e form a partition of unity, their gradients satisfy $\nabla \varphi_i^e = -\sum_{j \neq i} \nabla \varphi_j^e$. Using this property, we find that

$$\rho_{*,+}^e = \sum_{i=1}^N \max \left\{ 0, \sum_{\substack{j=1 \\ j \neq i}}^N (u_i^e - u_j^e) \int_{K^e} \varphi_i^e \mathbf{v} \cdot \nabla \varphi_j^e \mathrm{d}\mathbf{x} \right\} \leq \overline{\kappa} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N |u_i^e - u_j^e|,$$

where

$$\overline{\kappa} = \overline{\kappa}(e) := \max_{\substack{1 \leq i,j \leq N \\ i \neq j}} \left| \int_{K^e} \varphi_i^e \mathbf{v} \cdot \nabla \varphi_j^e \mathrm{d}\mathbf{x} \right|.$$

Without loss of generality, we assume that $u_1^e \leq \ldots \leq u_N^e$. Using this local numbering convention, we obtain the estimate

$$\begin{aligned} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N |u_i^e - u_j^e| &= \sum_{i=1}^N \left(\sum_{j=1}^{i-1} (u_i^e - u_j^e) + \sum_{j=i+1}^N (u_j^e - u_i^e) \right) \\ &\leq \sum_{i=1}^N \left(\sum_{j=1}^{i-1} (u_{\max}^e - u_j^e) + \sum_{j=i+1}^N (u_{\max}^e - u_i^e) \right) \\ \left(\sum_{i=1}^N \sum_{j=1}^{i-1} u_j^e = \sum_{i=1}^N (N-i) u_i^e \right) \quad &= \sum_{i=1}^N \left((N-1) u_{\max}^e - (N-i) u_i^e - (N-i) u_i^e \right) \end{aligned}$$

$$\left(\sum_{i=1}^{N}(N-1) = 2\sum_{i=1}^{N}(N-i)\right) = \sum_{i=1}^{N}(2(N-i)(u_{\max}^e - u_i^e))$$
$$\leq 2(N-1)\sum_{i=1}^{N}(u_{\max}^e - u_i^e),$$

which implies

$$\rho_{*,+}^e \leq 2\overline{\kappa}(N-1)\sum_{i=1}^{N}(u_{\max}^e - u_i^e).$$

A similar argument for the negative fluctuation yields

$$\rho_{*,-}^e \geq 2\overline{\kappa}(N-1)\sum_{i=1}^{N}(u_{\min}^e - u_i^e),$$

which proves the theorem. □

Remark 3. The PSI-corrected Rusanov scheme [5, 7] can be analyzed similarly. The corresponding element contributions $\rho_i^{e,L}$ are defined by (43) with $\Phi_{i,+}^e = \max\{0, \tilde{\rho}_i^{e,L}\}$ and $\Phi_{i,-}^e = \min\{0, \tilde{\rho}_i^{e,L}\}$. The LED property of the Rusanov weights $\tilde{\rho}_i^{e,L}$ implies the existence of bounded coefficients $\tilde{\kappa}_{i,\pm}^e \geq 0$ such that $\Phi_{i,+}^e = \tilde{\kappa}_{i,+}^e(u_{\max}^e - u_i^e)$ and $\Phi_{i,-}^e = \tilde{\kappa}_{i,-}^e(u_{\min}^e - u_i^e)$. Our definition (45) of the weights $\Phi_{i,\pm}^e$ bypasses the computation of $\tilde{\rho}_i^{e,L}$ by using $\tilde{\kappa}_{i,\pm}^e := 1$.

6. Subcell residual distribution

Following the recent trend toward the use of subcell approximations in bound-preserving schemes derived from high-order finite element discretizations of conservation laws [7, 21, 34, 37], we localize the RD procedure of Section 5 in a way which leads to optimal nonlinear LED schemes with compact stencils. Instead of distributing the fluctuations $\rho_{*,\pm}^e$ among the nodes of element K^e directly, we send them to subcells $K^{e,m}$, $m = 1, \ldots, M$ and perform residual distribution at the subcell level.

The natural subcell decomposition of a high-order Bernstein element is defined by its *Bézier net* [24], i.e., by a submesh whose vertices are located at the nodal points $\mathbf{x}_i^e$, $i = 1, \ldots, N$. The numbers of subcells containing $\mathbf{x}_i^e$ are stored in the integer set $\mathcal{S}_i^e$. The subset of $\{1, \ldots, N\}$ containing the element-level numbers of all nodes belonging to a subcell $K^{e,m}$ is denoted by $\mathcal{N}^{e,m}$. An example illustrating the subcell notation is shown in Fig. 1.

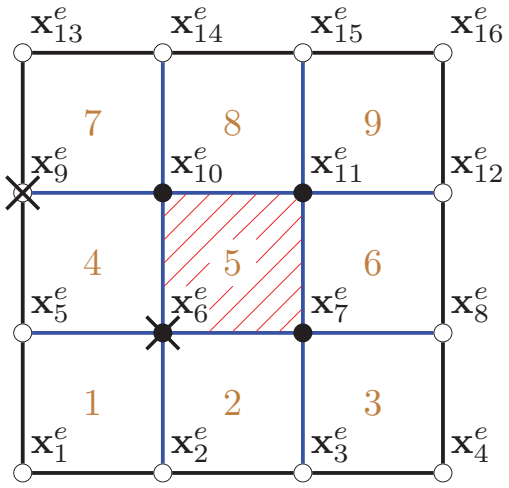

Fig. 1: Bernstein finite element approximation of polynomial degree $p = 3$ on a two-dimensional element K^e. The nodal points $\mathbf{x}^e_i,\ i = 1, \ldots, N = (p+1)^2 = 16$ and subcells $K^{e,m},\ m = 1, \ldots, M = p^2 = 9$ are numbered in lexicographical order but any other numbering may be used instead. Boundary and interior edges of the Bézier net are shown as black and blue line segments, respectively. Boundary and interior nodes are shown as white and black bullets, respectively. The host subcell sets of the two nodes marked by black crosses are $\mathcal{S}^e_6 = \{1, 2, 4, 5\}$ and $\mathcal{S}^e_9 = \{4, 7\}$. Element-level numbers of nodes belonging to the shaded subcell $K^{e,5}$ are stored in the set $\mathcal{N}^{e,5} = \{6, 7, 10, 11\}$.

The process of localized residual distribution begins with the definition of subcell fluctuations $\rho^{e,m}_{*,\pm}$ such that $\rho^e_{*,\pm} = \sum_{m=1}^M \rho^{e,m}_{*,\pm}$. The share of each cell is proportional to the fluctuation $\tilde{\rho}^{e,m}_{*,\pm}$ corresponding to the (multi-)linear interpolant $\tilde{u}^{e,m}_h$ of the Bernstein coefficients $u^e_j,\ j \in \mathcal{N}^{e,m}$. Introducing

$$\tilde{\rho}^{e,m}_{*,+} = \max\left\{0, -\int_{K^{e,m}} \mathbf{v}\cdot\nabla\tilde{u}^{e,m}_h \mathrm{d}\mathbf{x}\right\}, \qquad \tilde{\rho}^e_{*,+} = \sum_{m=1}^M \tilde{\rho}^{e,m}_{*,+},$$

$$\tilde{\rho}^{e,m}_{*,-} = \min\left\{0, -\int_{K^{e,m}} \mathbf{v}\cdot\nabla\tilde{u}^{e,m}_h \mathrm{d}\mathbf{x}\right\}, \qquad \tilde{\rho}^e_{*,-} = \sum_{m=1}^M \tilde{\rho}^{e,m}_{*,-},$$

we decompose the fluctuations $\rho^e_{*,\pm}$ into subcell contributions as follows:

$$\rho^{e,m}_{*,\pm} = \gamma^e_\pm \tilde{\rho}^{e,m}_{*,\pm}, \qquad \gamma^e_\pm = \frac{\rho^e_{*,\pm}}{\tilde{\rho}^e_{*,\pm} \pm \epsilon}. \tag{49}$$

This distribution procedure uses the Bézier net fluctuations $\tilde{\rho}^{e,m}_{*,\pm}$ as nonconservative *targets* which are multiplied by $\gamma^e_\pm$ to enforce conservation.

Next, the subcell shares $\rho_{*,\pm}^{e,m}$ are distributed among the nodes of $K^{e,m}$ as in Section 5 using localized LED-type weights $\Phi_{i,\pm}^{e,m}$ to define

$$\omega_{i,+}^{e,m} = \begin{cases} \dfrac{\Phi_{i,+}^{e,m}}{\Phi_{+}^{e,m} + \epsilon} & \text{if } i \in \mathcal{N}^{e,m}, \\ 0 & \text{otherwise,} \end{cases} \qquad \Phi_{+}^{e,m} = \sum_{j \in \mathcal{N}^{e,m}} \Phi_{j,+}^{e,m}, \tag{50}$$

$$\omega_{i,-}^{e,m} = \begin{cases} \dfrac{\Phi_{i,-}^{e,m}}{\Phi_{-}^{e,m} - \epsilon} & \text{if } i \in \mathcal{N}^{e,m}, \\ 0 & \text{otherwise,} \end{cases} \qquad \Phi_{-}^{e,m} = \sum_{j \in \mathcal{N}^{e,m}} \Phi_{j,-}^{e,m} \tag{51}$$

and

$$\rho_i^{e,m,L} = \rho_{*,+}^{e,m}\omega_{i,+}^{e,m} + \rho_{*,-}^{e,m}\omega_{i,-}^{e,m}, \qquad i = 1, \dots, N. \tag{52}$$

For example, a subcell version of the PSI-corrected Rusanov scheme (as defined in Section 5) can be constructed using the Bézier net weights

$$\Phi_{i,+}^{e,m} = \max\left\{0, -\int_{K^{e,m}} \tilde{\varphi}_i^e \mathbf{v} \cdot \nabla \tilde{u}_h^{e,m} \mathrm{d}\mathbf{x} + d^{e,m}|\mathcal{N}^{e,m}|(\bar{u}^{e,m} - u_i^e)\right\}, \tag{53}$$

$$\Phi_{i,-}^{e,m} = \min\left\{0, -\int_{K^{e,m}} \tilde{\varphi}_i^e \mathbf{v} \cdot \nabla \tilde{u}_h^{e,m} \mathrm{d}\mathbf{x} + d^{e,m}|\mathcal{N}^{e,m}|(\bar{u}^{e,m} - u_i^e)\right\}, \tag{54}$$

where $\tilde{\varphi}_i^e$, $i \in \mathcal{N}^{e,m}$ is the Lagrange basis function of the $p = 1$ subcell approximation, $|\mathcal{N}^{e,m}|$ is the number of DOFs per subcell, $\bar{u}^{e,m} = \frac{1}{|\mathcal{N}^{e,m}|}\sum_{j \in \mathcal{N}^{e,m}} u_j^e$ is the subcell average of the Bernstein coefficients, and

$$d^{e,m} \geq \left| \int_{K^{e,m}} \tilde{\varphi}_i^e \mathbf{v} \cdot \nabla \tilde{\varphi}_j^e \mathrm{d}\mathbf{x} \right| \qquad \forall i, j \in \mathcal{N}^{e,m}$$

is the artificial diffusion coefficient of the subcell contribution. Although the estimation of $d^{e,m}$ is as cheap and easy as that of d^e in the case $p = 1$, it is hardly worth the effort. For reasons explained in Section 5, we prefer to bypass the computation of $d^{e,m}$ and replace (53),(54) with

$$\Phi_{i,+}^{e,m} = u_{\max}^{e,m} - u_i^e, \qquad u_{\max}^{e,m} = \max_{j \in \mathcal{N}^{e,m}} u_j^e, \tag{55}$$

$$\Phi_{i,-}^{e,m} = u_{\min}^{e,m} - u_i^e, \qquad u_{\min}^{e,m} = \min_{j \in \mathcal{N}^{e,m}} u_j^e. \tag{56}$$

Summing over $m \in \mathcal{S}_i^e$, we obtain the element contributions

$$\rho_i^{e,L} = \sum_{m \in \mathcal{S}_i^e} \rho_i^{e,m,L} = \gamma_+^e \tilde{\rho}_{i,+}^{e,L} + \gamma_-^e \tilde{\rho}_{i,-}^{e,L}, \qquad i = 1, \dots, N, \tag{57}$$

where

$$\tilde{\rho}_{i,+}^{e,L} = \sum_{m=1}^{M} \tilde{\rho}_{*,+}^{e,m} \omega_{i,+}^{e,m} = \sum_{m \in \mathcal{S}_i^e} \frac{\tilde{\rho}_{*,+}^{e,m} (u_{\max}^{e,m} - u_i^e)}{\sum_{j \in \mathcal{N}^{e,m}} (u_{\max}^{e,m} - u_j^e) + \epsilon}, \tag{58}$$

$$\tilde{\rho}_{i,-}^{e,L} = \sum_{m=1}^{M} \tilde{\rho}_{*,-}^{e,m} \omega_{i,-}^{e,m} = \sum_{m \in \mathcal{S}_i^e} \frac{\tilde{\rho}_{*,-}^{e,m} (u_{\min}^{e,m} - u_i^e)}{\sum_{j \in \mathcal{N}^{e,m}} (u_{\min}^{e,m} - u_j^e) - \epsilon}. \tag{59}$$

The LED property of $\tilde{\rho}_{i,\pm}^{e,L}$ can be shown following the proof of Theorem 4. The element contributions defined by (57) inherit it provided that the coefficients $\gamma_\pm^e$ are bounded. In the limit $\epsilon \searrow 0$, which is tacitly implied in all formulas involving ϵ, the conservation property follows from the fact that

$$\sum_{i=1}^{N} \tilde{\rho}_{i,\pm}^{e,L} = \sum_{i=1}^{N} \sum_{m=1}^{M} \tilde{\rho}_{*,\pm}^{e,m} \omega_{i,+}^{e,m} = \sum_{m=1}^{M} \tilde{\rho}_{*,\pm}^{e,m} \sum_{i=1}^{N} \omega_{i,+}^{e,m} = \tilde{\rho}_{*,\pm}^{e}, \tag{60}$$

which implies

$$\sum_{i=1}^{N} \rho_i^{e,L} = \sum_{i=1}^{N} \left(\gamma_+^e \tilde{\rho}_{i,+}^{e,L} + \gamma_-^e \tilde{\rho}_{i,-}^{e,L} \right) = \rho_{*,+}^e + \rho_{*,-}^e = \rho_*^e.$$

The subcell RD scheme defined by (57)–(59) is far more accurate than the element-stencil version presented in Section 5. However, it has the theoretical disadvantage that the LED property cannot be guaranteed under realistic time step restrictions if the coefficients $\gamma_\pm^e$ become unbounded or very large.

While violations of local bounds are very unlikely to occur in practice, we propose a way to perform the necessary correction if this turns out to be the case. The idea is to impose an upper bound $\gamma \gg 1$ on $\gamma_\pm^e$ and redistribute the remainders $(\gamma_\pm^e - \min\{\gamma, \gamma_\pm^e\})\tilde{\rho}_{*,\pm}^e$ among the nodes of K^e using the element-stencil weights (45) to ensure boundedness. The so-defined *failsafe version* of the subcell RD scheme replaces (57) with

$$\begin{aligned} \rho_i^{e,L} = {} & \min\{\gamma, \gamma_+^e\} \tilde{\rho}_{i,+}^{e,L} + \frac{(\gamma_+^e - \min\{\gamma, \gamma_+^e\}) \tilde{\rho}_{*,+}^e}{\sum_{j=1}^{N} (u_{\max}^e - u_j^e) + \epsilon} (u_{\max}^e - u_i^e) \\ & + \min\{\gamma, \gamma_-^e\} \tilde{\rho}_{i,-}^{e,L} + \frac{(\gamma_-^e - \min\{\gamma, \gamma_-^e\}) \tilde{\rho}_{*,-}^e}{\sum_{j=1}^{N} (u_{\min}^e - u_j^e) - \epsilon} (u_{\min}^e - u_i^e). \end{aligned} \tag{61}$$

If the coefficients $\gamma_{\pm}^e$ become larger than γ, the unacceptable fluctuation is distributed using the element-stencil formula which guarantees the LED property by Theorem 4. Since the high-order Bernstein polynomial u_h^e is at least as smooth as its piecewise-multilinear Bézier net approximation $\tilde{u}_h^e$ [24], situations in which the magnitude of $\rho_{*,\pm}^e$ is much greater than that of $\tilde{\rho}_{*,\pm}^e$ are rare. Following the design philosophy of *a posteriori* limiting [19, 21], modification (61) may be used to recalculate the solution in the unlikely case in which definition (57) does produce an undershoot or overshoot.

Remark 4. Representations (57) and (61) of the subcell RD schemes are convenient for theoretical analysis purposes. In practice, element contributions to each node should be assembled from subcell contributions in the same way in which global vectors are assembled from element vectors in CG methods. In the case $\gamma_{\pm}^e > \gamma$, the numerically stable representation $\rho_{*,\pm}^e - \min\{\gamma\tilde{\rho}_{*,\pm}^e, \rho_{*,\pm}^e\}$ of $(\gamma_{\pm}^e - \min\{\gamma, \gamma_{\pm}^e\})\tilde{\rho}_{*,\pm}^e$ or the if-definition of the implied limit $\epsilon \searrow 0$ must be employed in (61) to avoid potentially significant conservation errors due to the presence of $\epsilon > 0$ in the denominators of $\gamma_{\pm}^e$.

In addition to the LED structure of $\tilde{\rho}_{i,\pm}^{e,L}$ and preservation of $\rho_{*,\pm}^e$, the failsafe subcell RD scheme (61) guarantees the LED property of $\rho_{i,\pm}^{e,L}$.

Theorem 5. *There exists a constant $\kappa = \kappa(e, N, \gamma, d) > 0$ such that the RD scheme defined by (61) is LED with bounded coefficients $\hat{\kappa}_{i,\pm}^e \le \kappa$.*

Proof. Adapting the proof of Theorem 4 to subcells, it is easy to verify that

$$\tilde{\rho}_{i,+}^{e,L} = \sum_{m\in\mathcal{S}_i^e} \frac{\tilde{\rho}_{*,+}^{e,m}(u_{\max}^{e,m} - u_i^e)}{\sum_{j\in\mathcal{N}^{e,m}}(u_{\max}^{e,m} - u_j^e) + \epsilon} \le \kappa_i \sum_{m\in\mathcal{S}_i^e} (u_{\max}^{e,m} - u_i^e) \le \kappa_i |\mathcal{S}_i^e| (u_{\max}^e - u_i^e),$$

where $|\mathcal{S}_i^e|$ is the number of subcells containing the point $\mathbf{x}_i^e$. Similarly, we have $\tilde{\rho}_{i,-}^{e,L} \ge \kappa_i |\mathcal{S}_i^e| (u_{\min}^e - u_i^e)$ for some bounded $\kappa_i = \kappa_i(e) > 0$.

Invoking the definition of $\gamma_{\pm}^e$, we consider the cases $\gamma_{\pm}^e \le \gamma$ and $\gamma < \gamma_{\pm}^e$ separately. In both cases, there exist scaling factors $\beta_{\pm}^e \in [0, 1]$ such that

$$(\gamma_{\pm}^e - \min\{\gamma, \gamma_{\pm}^e\})\tilde{\rho}_{*,\pm}^e = \beta_{\pm}^e \rho_{*,\pm}^e.$$

Applying Theorem 4 with constant $\overline{\kappa}$ to element contributions, we obtain

$$\begin{aligned}\min\{\gamma,\gamma_+^e\}\tilde{\rho}_{i,+}^{e,L} &+ \frac{(\gamma_+^e-\min\{\gamma,\gamma_+^e\})\tilde{\rho}_{*,+}^e}{\sum_{j=1}^N(u_{\max}^e-u_j^e)+\epsilon}(u_{\max}^e-u_i^e)\\ &\le\left(\gamma\kappa_i|\mathcal{S}_i^e|+\frac{\beta_+^e\rho_{*,+}^e}{\sum_{j=1}^N(u_{\max}^e-u_j^e)+\epsilon}\right)(u_{\max}^e-u_i^e)\\ &\le(\gamma\kappa_i|\mathcal{S}_i^e|+\overline{\kappa})(u_{\max}^e-u_i^e),\end{aligned}$$

because $\beta_\pm^e\le 1$. A similar estimate for the negative part yields

$$\begin{aligned}\min\{\gamma,\gamma_-^e\}\tilde{\rho}_{i,-}^{e,L} &+ \frac{(\gamma_-^e-\min\{\gamma,\gamma_-^e\})\tilde{\rho}_{*,-}^e}{\sum_{j=1}^N(u_{\min}^e-u_j^e)-\epsilon}(u_{\min}^e-u_i^e)\\ &\ge(\gamma\kappa_i|\mathcal{S}_i^e|+\overline{\kappa})(u_{\min}^e-u_i^e).\end{aligned}$$

This proves the existence of constants $0\le\hat{\kappa}_{i,\pm}^e\le(\gamma\kappa_i|\mathcal{S}_i^e|+\overline{\kappa})$ such that

$$\rho_i^{e,L}=\hat{\kappa}_{i,+}^e(u_{\max}^e-u_i^e)+\hat{\kappa}_{i,-}^e(u_{\min}^e-u_i^e).$$

□

For $\gamma\ge\max\{\gamma_{i,+}^e,\gamma_{i,-}^e\}$, the γ-controlled subcell RD scheme (61) reduces to the basic version (57). In this case, the LED property of $\rho_i^{e,L}$ can be shown with respect to the tightened subcell-stencil bounds $\tilde{u}_{i,\min}=\min_{m\in\mathcal{S}_i}u_{\min}^{e,m}$ and $\tilde{u}_{i,\max}=\max_{m\in\mathcal{S}_i}u_{\max}^{e,m}$. All numerical results to be presented in this work were obtained with $\gamma=10$ because this value was found to satisfy the LED criterion for approximately the same CFL numbers as linear low-order schemes. The choice $\gamma=1$ produces more diffusive results because a smaller fraction of the total fluctuation can be distributed using the localized subcell weights. The use of $\gamma=100$ leads to violations of the CFL condition (21) if the same time step is employed. To compensate the tenfold increase in the value of γ and rule out formation of spurious oscillations, the time step should be refined by a factor of 10 in accordance with the proof of Theorem 5 in which linear dependence of the constant κ on γ was shown. The sharp upper bound for time steps satisfying (21) can be calculated using κ_i defined by (22). If it turns out to be impractically small for some nodes, a smaller value of γ may be employed in elements containing these nodes.

7. High-order FCT-based schemes

For any $p \in \mathbb{N}$, the accuracy of the nonlinear subcell RD scheme presented in Section 6 is similar to that of the matrix-based discrete upwinding method [12, 29] for a (multi-)linear finite element approximation on the submesh with the same nodal points. To achieve optimal convergence behavior for high-order Bernstein finite element discretizations, we invoke Theorem 3 and perform bound-preserving antidiffusive corrections using the element-based FCT algorithms developed for DG and CG methods in [8, 34].

The linear systems of the high- and low-order schemes read

$$\sum_{j=1}^{N_h} m_{ij} u_j^H = \sum_{j=1}^{N_h} m_{ij} u_j + \Delta t \sum_{e \in \mathcal{E}_i} (\rho_i^{e,H} + \sigma_i^{e,H}), \qquad i = 1, \ldots, N_h, \tag{62}$$

$$m_i u_i^L = m_i u_i + \Delta t \sum_{e \in \mathcal{E}_i} (\rho_i^{e,L} + \sigma_i^{e,L}), \qquad i = 1, \ldots, N_h. \tag{63}$$

Introducing the high-order Galerkin approximations

$$\dot{u}_i^H = \frac{u_i^H - u_i}{\Delta t} \approx \frac{\mathrm{d}u_i}{\mathrm{d}t}$$

to the time derivatives of global DOFs, system (62) can be written as

$$\begin{aligned} m_i u_i^H &= m_i u_i + \sum_{j=1}^{N_h} (m_i \delta_{ij} - m_{ij})(u_j^H - u_j) + \Delta t \sum_{e \in \mathcal{E}_i} (\rho_i^{e,H} + \sigma_i^{e,H}) \\ &= m_i u_i + \Delta t \sum_{j=1}^{N_h} (m_i \delta_{ij} - m_{ij}) \dot{u}_j^H + \Delta t \sum_{e \in \mathcal{E}_i} (\rho_i^{e,H} + \sigma_i^{e,H}) \\ &= m_i u_i^L - \Delta t \sum_{e \in \mathcal{E}_i} (\rho_i^{e,L} + \sigma_i^{e,L}) + \Delta t \sum_{e \in \mathcal{E}_i} \sum_{j=1}^{N} (m_i^e \delta_{ij} - m_{ij}^e) \dot{u}_j^{e,H} \\ &\quad + \Delta t \sum_{e \in \mathcal{E}_i} (\rho_i^{e,H} + \sigma_i^{e,H}) = m_i u_i^L + \Delta t \sum_{e \in \mathcal{E}_i} \eta_i^{e,H}, \end{aligned} \tag{64}$$

where

$$\eta_i^{e,H} = (\rho_i^{e,H} + \sigma_i^{e,H}) - (\rho_i^{e,L} + \sigma_i^{e,L}) + \sum_{j=1}^{N} (m_i^e \delta_{ij} - m_{ij}^e) \dot{u}_j^{e,H} \tag{65}$$

are raw antidiffusive element contributions. Note that $\sum_{j=1}^N \eta_j^{e,H} = 0$ in view of the requirement (31) and the fact that $m_i^e = \sum_{j=1}^N m_{ij}^e$ by definition.

As shown in the proof of Theorem 3, it is sufficient to impose local bounds on the one-sided approximations $\bar{u}_i^e$ to the Bernstein coefficients. In the CG version, the bound-preserving final value is the convex average

$$\bar{u}_i = \frac{1}{m_i} \sum_{e \in \mathcal{E}_i} m_i^e \bar{u}_i^e, \qquad i = 1, \ldots, N_h. \tag{66}$$

The high- and low-order DG approximations to $\bar{u}_i^e$ are given by

$$m_i^e u_i^{e,H} = m_i^e u_i^{e,L} + \Delta t \eta_i^{e,H}, \tag{67}$$

$$m_i^e u_i^{e,L} = m_i^e u_i^e + \Delta t (\rho_i^{e,L} + \sigma_i^{e,L}). \tag{68}$$

In the CG version, the low-order approximation $u_i^{e,L}$ defined by (68) may be replaced with the solution $u_i^L = \frac{1}{m_i} \sum_{e \in \mathcal{E}_i} m_i^e u_i^{e,L}$ of (63). Both versions produce the solution $u_i^H = \frac{1}{m_i} \sum_{e \in \mathcal{E}_i} m_i^e u_i^{e,H}$ of (64) after global assembly.

Note that $\eta_i^{e,H}$ can be calculated in a matrix-free manner but computation of the nodal time derivatives that appear in (65) requires solution of linear systems with the mass matrix M_C. In the DG version, this matrix is block-diagonal and, therefore, $\dot{u}^{e,H}$ can be obtained by solving the local problem

$$M_C^e \dot{u}^{e,H} = \rho^{e,H} + \sigma^{e,H}. \tag{69}$$

An algorithm for fast inversion of the simplicial Bernstein mass matrix in DG methods can be found in [26]. Another way to calculate $\dot{u}^{e,H}$ efficiently in CG and DG methods is based on the truncated series approximation [25]

$$M_C^{-1} \approx M_L^{-1} \sum_{k=0}^{m} (I - M_C M_L^{-1})^k, \qquad m \geq 0. \tag{70}$$

For practical purposes, it is sufficient to use $m = 1$, which yields

$$\dot{u}^H = \left[I + M_L^{-1} (M_L - M_C) \right] M_L^{-1} \left(\rho^H + \sigma^H \right). \tag{71}$$

This definition makes it possible to calculate $\dot{u}^H$ in a matrix-free manner. Related approaches to avoiding inversion of consistent mass matrices in explicit Runge-Kutta time integrators for CG schemes can be found in [3].

To satisfy the LED condition (30) for $\tilde{u}_i^e = u_i^{e,L}$, we constrain the antidiffusive element contributions as follows [8, 34]:

1. Replace $\eta_i^{e,H}$ by the bound-preserving nonconservative approximation

$$\bar{\eta}_i^{e,H} = \frac{m_i^e}{\Delta t}\left(\bar{u}_i^{e,H} - u_i^{e,L}\right), \tag{72}$$

$$\bar{u}_i^{e,H} = \min\left\{u_{i,\max}, \max\left\{u_i^{e,H}, u_{i,\min}\right\}\right\}. \tag{73}$$

2. Perform additional limiting to enforce the conservation property

$$\eta_i^e = \begin{cases} -\dfrac{\bar{\eta}_{*,-}^e}{\bar{\eta}_{*,+}^e}\bar{\eta}_i^{e,H} & \text{if} \quad \bar{\eta}_i^{e,H} > 0 \ \wedge \ \bar{\eta}_{*,+}^e + \bar{\eta}_{*,-}^e > 0, \\ -\dfrac{\bar{\eta}_{*,+}^e}{\bar{\eta}_{*,-}^e}\bar{\eta}_i^{e,H} & \text{if} \quad \bar{\eta}_i^{e,H} < 0 \ \wedge \ \bar{\eta}_{*,+}^e + \bar{\eta}_{*,-}^e < 0, \\ \bar{\eta}_i^{e,H} & \text{otherwise,} \end{cases} \tag{74}$$

$$\bar{\eta}_{*,+}^e = \sum_{j=1}^N \max\left\{0, \bar{\eta}_j^{e,H}\right\}, \qquad \bar{\eta}_{*,-}^e = \sum_{j=1}^N \min\left\{0, \bar{\eta}_j^{e,H}\right\}. \tag{75}$$

This two-step correction procedure can be interpreted as pointwise limiting of $\eta_i^{e,H}$ followed by residual distribution with weights defined in terms of $\bar{\eta}_j^{e,H}$.

Replacing $\eta_i^{e,H}$ by the limited antidiffusive element contributions η_i^e, the constrained DOFs can be calculated using the formula

$$m_i^e \bar{u}_i^e = m_i^e u_i^{e,L} + \Delta t \eta_i^e \tag{76}$$

in the DG version (in which $\bar{u}_i = \bar{u}_i^e$ is the final result) and

$$m_i \bar{u}_i = m_i u_i^L + \Delta t \sum_{e \in \mathcal{E}_i} \eta_i^e \tag{77}$$

in the CG version (in which $\bar{u}_i$ is a convex average of $\bar{u}_i^e$). For a detailed description of such FCT algorithms, we refer the reader to [8, 34].

8. Numerical examples

In this section, we illustrate the accuracy of the presented schemes by 1D, 2D, and 3D numerical examples for finite element approximations with Bernstein polynomials of degree $p \in \mathbb{N}$. A matter of particular interest is

the ability or inability of a given method to deliver p-independent rates of convergence with respect to the total number of DOFs (#DOFs). Therefore, many examples depict results obtained with different values of p for mesh sizes $h(p)$ corresponding to the same constant number of unknowns.

Let N_h^{CG} and N_h^{DG} denote #DOFs for the CG and DG version, respectively. In one dimension, there are $p+1$ nodes per element and E_h-1 shared DOFs in the CG case. It follows that $N_h^{\mathrm{CG}} = pE_h+1$ and $N_h^{\mathrm{DG}} = (p+1)E_h$ in 1D. In the multi-dimensional case, the values of N_h^{CG} and N_h^{DG} for different combinations of h and p on Cartesian meshes with the same resolution in each dimension can be determined similarly. For example, $N_h^{\mathrm{CG}} = 121^2 = (1 \cdot 120 + 1)^2 = (2 \cdot 60 + 1)^2 = (3 \cdot 40 + 1)^2$ may correspond to bilinear, biquadratic, or bicubic CG approximations ($p = 1, 2,$ or 3) on grids with sizes $h = \frac{1}{120}, \frac{1}{60},$ or $\frac{1}{40}$, respectively. In the DG case, the same values of N_h^{DG} are obtained for (p, h) and $(2p + 1, 2h)$. That is, a fair comparison of the DG results for p and $2p + 1$ requires the use of meshes that differ by one refinement level. This criterion makes it possible to keep N_h^{DG} fixed, e.g., for $p = 1, 3, 7, \ldots$ or $p = 2, 5, 11, \ldots$ by using mesh sizes $h, 2h, 4h, \ldots$.

The following definitions of $\rho_i^{e,L}$ are compared and evaluated in this study:

DU(p)	discrete upwinding defined by (34), (35);
PDU(p)	preconditioned discrete upwinding (36);
RU(p)	linear Rusanov scheme defined by (37)-(39);
PRUp)	PSI-corrected Rusanov scheme (cf. [1, 5, 36]);
RUS(p)	subcell Rusanov scheme defined by (50)–(54);
RD(p)	nonlinear RD scheme (43) with weights (45);
RDS(p)	subcell RD scheme (61) with weights (55).

The lumped-mass definition (33) of the boundary terms $\sigma_i^{e,L}$ is adopted in 2D and 3D implementations of all low-order schemes. In the 1D case, the element matrices S^e and $\tilde{S}^e$ coincide and, therefore, the Galerkin element contributions $\sigma_i^{e,H} = \sigma_i^{e,L}$ do not require any modifications.

The suffix FCT is appended to the above abbreviations if limited antidiffusive corrections of $\rho^{e,L} + \sigma^{e,L}$ are performed. All FCT schemes are based on the splitting (67),(68) and use algorithm (72)–(75) to calculate η_i^e.

In some 1D examples, we compare the CG and DG results for different methods. The corresponding finite element approximations of degree p are denoted by CG(p) and DG(p), respectively. Numerical studies of multidimensional problems are restricted to the DG version and conducted using

the open-source C++ finite element library MFEM [35].

In all examples, we integrate in time using the optimal explicit SSP Runge-Kutta methods of second order (in 1D) or third order (in 2D and 3D). The time step Δt is chosen to be sufficiently small for the temporal discretization error to be negligible and the CFL condition to be fulfilled.

8.1. Advection of smooth and discontinuous data in 1D

For a preliminary evaluation of different schemes, we solve the one-dimensional version of equation (1) with constant velocity $v = 1$ in $\Omega = (0, 1)$. In this set of numerical experiments, we advect the smooth cosine hill

$$u_0(x) = \begin{cases} \frac{1}{2}\left[1 + \cos\left(\pi \frac{x-0.25}{0.15}\right)\right] & \text{if } |x - 0.25| < 0.15, \\ 0, & \text{otherwise} \end{cases} \tag{78}$$

and the discontinuous step function

$$u_0(x) = \begin{cases} 1 & \text{if } |x - 0.25| < 0.15, \\ 0 & \text{otherwise.} \end{cases} \tag{79}$$

The inflow boundary condition for both test cases is $u(0, t) = 0,\ \forall t \geq 0$. The exact solution of the initial-boundary value problem is given by

$$u(x, t) = \begin{cases} u_0(x - vt) & \text{if } x > vt, \\ 0 & \text{otherwise.} \end{cases} \tag{80}$$

Computations are performed on uniform meshes of one-dimensional linear and cubic Bernstein finite elements such that $N_h^{\mathrm{CG}} = 121 = 1 \cdot 120 + 1 = 3 \cdot 40 + 1$ and $N_h^{\mathrm{DG}} = 120 = 2 \cdot 60 = 4 \cdot 30$. The results are presented in Fig. 2–3. In each diagram, we show the initial condition and numerical solutions at the final time $T = 0.5$ computed with $\Delta t = 10^{-3}$.

The relative performance of different approximations is similar in both examples. Basic discrete upwinding is inferior to the preconditioned version because only the latter converges independently of p. Note that the DU(1), PDU(1), and PDU(3) results are virtually indistinguishable, whereas the DU(3) solution is less accurate. The same qualitative behavior is observed for the RU/RUS and RD/RDS pairs of low-order schemes. As shown in Fig. 4, the RU(3) approximation is significantly more diffusive than the RU(1) result. The PSI correction in PRU(3) leads to a marked improvement, whereas RUS(3) performs almost as well as RU(1). The accuracy of RDS(3)

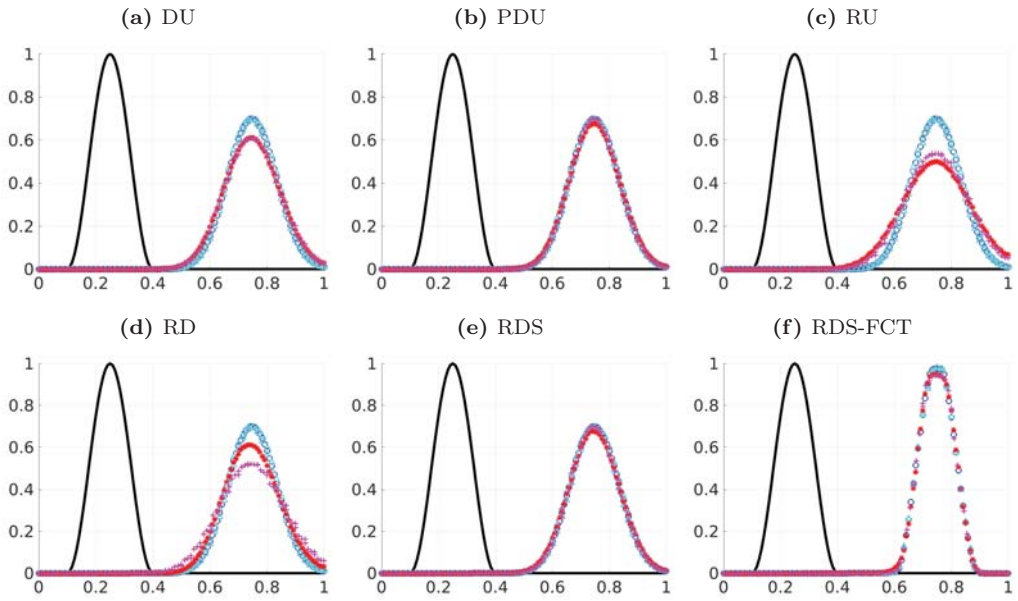

Fig. 2: Advection of a smooth profile in 1D. Initial condition: black line, CG(1): dark blue circles, DG(1): bright blue triangles, CG(3): red stars, DG(3): pink x-marks.

is superior to that of RD(3) and comparable to that of PDU(3). Among the low-order schemes, PDU, RUS, and RDS produce the best results for large p and fixed #DOFs. The FCT correction step brings about a further dramatic improvement, as demonstrated by the RDS-FCT results in Fig. 2f and 3f. The initial shapes of the advected profiles are captured very well, and the p-independent convergence behavior is largely preserved. Remarkably, there is hardly any difference between the CG and DG results for $N_h^{\mathrm{CG}} \approx N_h^{\mathrm{DG}}$.

As a preliminary conclusion that can be drawn from these 1D tests, we recommend the use of PDU or RDS as low-order methods for $p > 1$ if the Galerkin element matrices are readily available and RDS for matrix-free alternatives. In what follows, we focus on numerical studies of DG-based PDU and RDS approximations in 2D and 3D. The resolving power of these schemes is illustrated by comparisons to DU results. Since the basic RD scheme was found to be quite diffusive even in 1D, we show the RD results in just one 2D example. The results obtained with Rusanov weights (i.e., using RU, PRU, or RUS) are omitted because even RUS becomes less accurate than RDS while the involved computational effort is higher, as pointed out in Section 6.

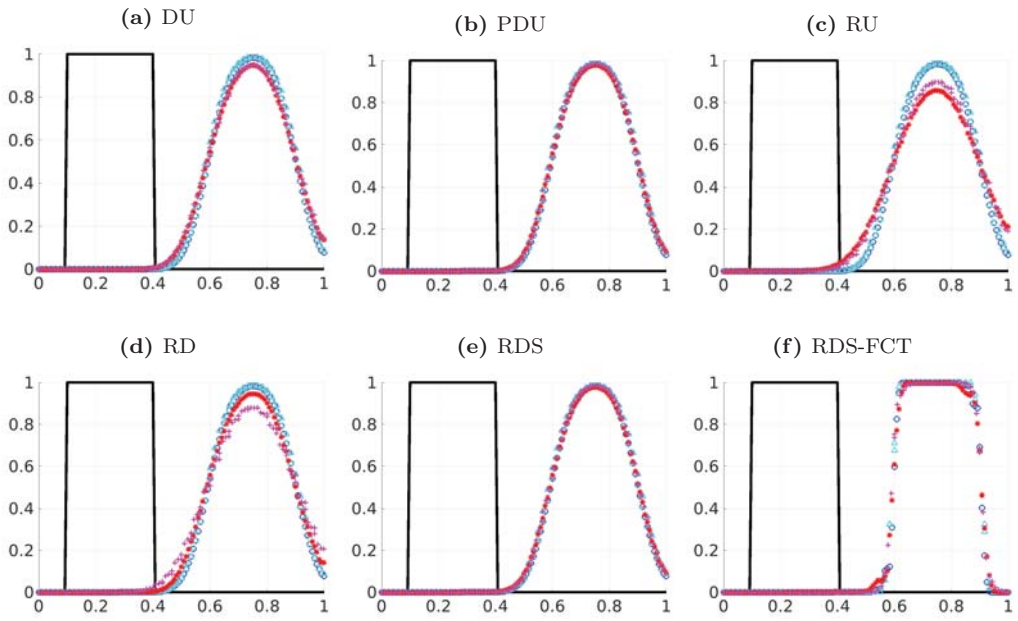

Fig. 3: Advection of a discontinuous profile in 1D. Initial condition: black line, CG(1): dark blue circles, DG(1): bright blue triangles, CG(3): red stars, DG(3): pink x-marks.

8.2. Solid body rotation in 2D

In this standard 2D test [32], we use $\mathbf{v}(x,y) = (0.5-y, x-0.5)^T$ to rotate a slotted cylinder, a sharp cone, and a smooth hump around the center of $\Omega = (0,1)^2$. The initial condition, as shown in Fig. 5a and 5d, is given by

$$u_0(x,y) = \begin{cases} u_0^{\text{hump}}(x,y) & \text{if } \sqrt{(x-0.25)^2+(y-0.5)^2} \leq 0.15, \\ u_0^{\text{cone}}(x,y) & \text{if } \sqrt{(x-0.5)^2+(y-0.25)^2} \leq 0.15, \\ 1 & \text{if } \begin{cases} \left(\sqrt{(x-0.5)^2+(y-0.75)^2} \leq 0.15\right) \wedge \\ \left(|x-0.5| \geq 0.025 \vee \ y \geq 0.85\right), \end{cases} \\ 0 & \text{otherwise,} \end{cases}$$

where

$$u_0^{\text{hump}}(x,y) = \frac{1}{4} + \frac{1}{4}\cos\left(\frac{\pi\sqrt{(x-0.25)^2+(y-0.5)^2}}{0.15}\right), \tag{81}$$

$$u_0^{\text{cone}}(x,y) = 1 - \frac{\sqrt{(x-0.5)^2+(y-0.25)^2}}{0.15}. \tag{82}$$

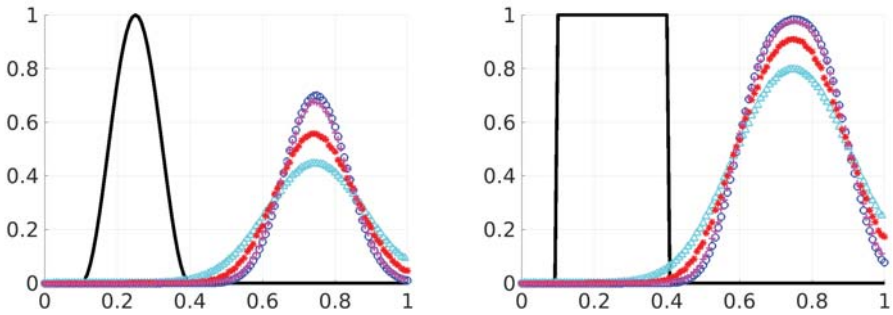

Fig. 4: Rusanov-based CG results for 1D advection of a smooth (left panel) and discontinuous (right panel) profile. Initial condition: black line, RU(1): dark blue circles, RU(3): bright blue triangles, PRU(3): red stars, RUS(3): pink x-marks.

Homogeneous Dirichlet boundary conditions are prescribed at the inlets.

After each full rotation, the exact solution $u(x, y, 2\pi k),\ k \in \mathbb{N}$ coincides with $u_0(x, y)$. The challenge of this test is to preserve the shape of the projected initial condition $u_h(\cdot, 0)$ as accurately as possible. We present the DG solutions at the final time $T = 2\pi$ (i.e., after one full rotation) obtained with $N_h^{\mathrm{DG}} = 144^2$ corresponding to $p \in \{2, 5\}$ for all schemes under investigation and, additionally, $p = 11$ for RDS and RDS-FCT. The time step $\Delta t = \frac{\pi}{2} \cdot 10^{-3}$ was used in all simulations for this test problem.

The snapshots of numerical solutions in Fig. 5 and slices displayed in Fig. 6 confirm that PDU is more accurate than DU. However, in contrast to the 1D advection with constant velocity, PDU becomes slightly more diffusive as p increases. We investigate this aspect further in Section 8.3. The matrix-free RDS scheme outperforms both matrix-based approaches, and its accuracy is largely independent of p. The strange-looking accumulation of mass to the right of the hump in Fig. 6g–6i is caused by diffusive fluxes that transport the remainders of the cylinder in all directions, see Fig. 5b–5c,5e–5i. The slices of the RDS-FCT solutions are not affected by these fluxes because their magnitude is significantly reduced in the process of antidiffusive corrections, cf. Fig. 5j–5l. The cuts through the slotted cylinder, as shown in Fig. 6j–6l, illustrate the shock-capturing capabilities of different methods and the destructive effect of numerical diffusion. All low-order schemes convert the cylinder into a smooth hump, whereas RDS-FCT preserves the slot fairly well. A comparison of the local minima inside the slot reveals that the FCT results become more diffusive as p increases. This behavior reflects the

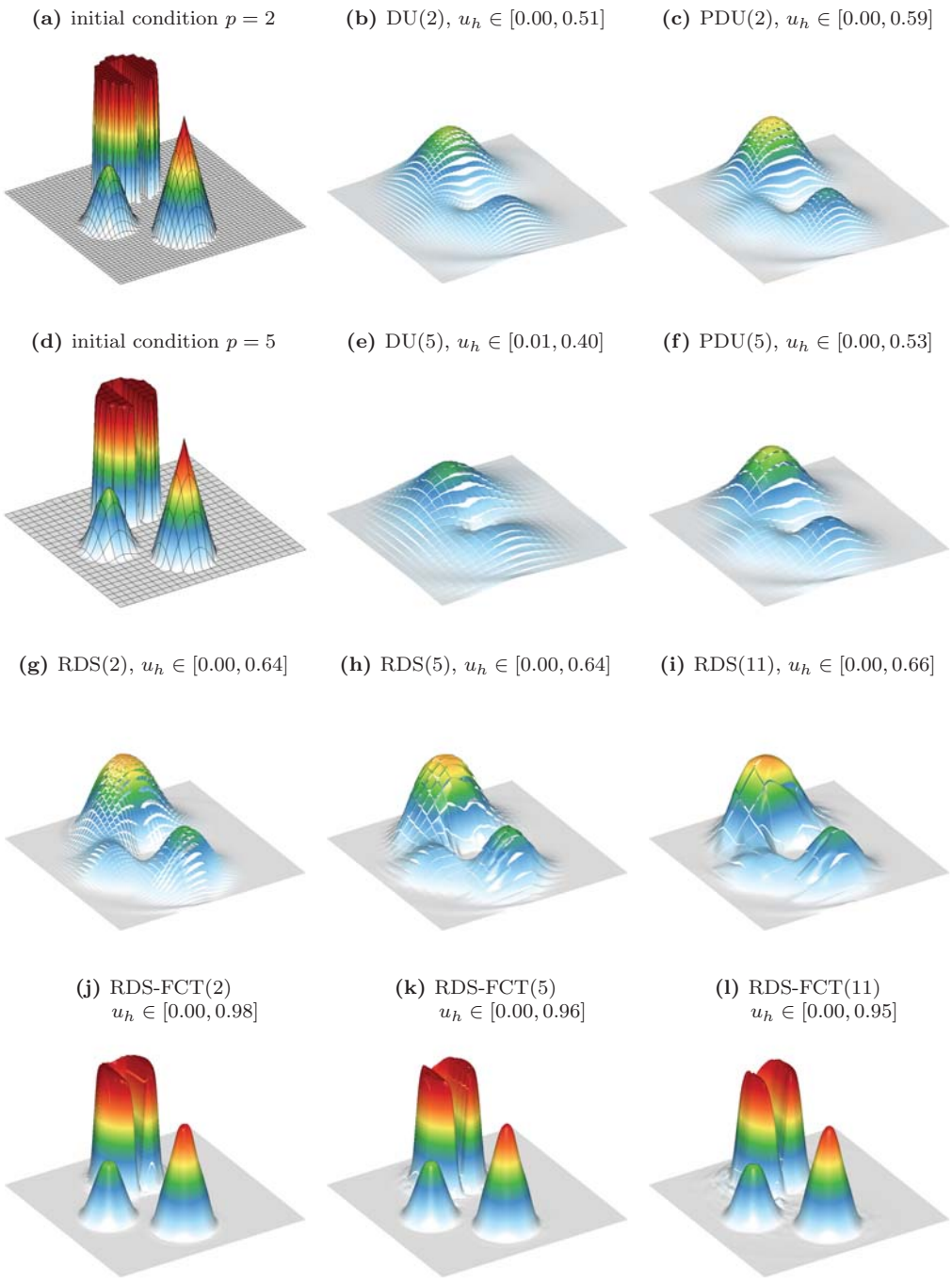

Fig. 5: Solid body rotation in 2D. Initial condition and numerical solutions $u_h(\cdot, 2\pi)$ obtained using low-order schemes and FCT for $p \in \{2, 5, 11\}$.

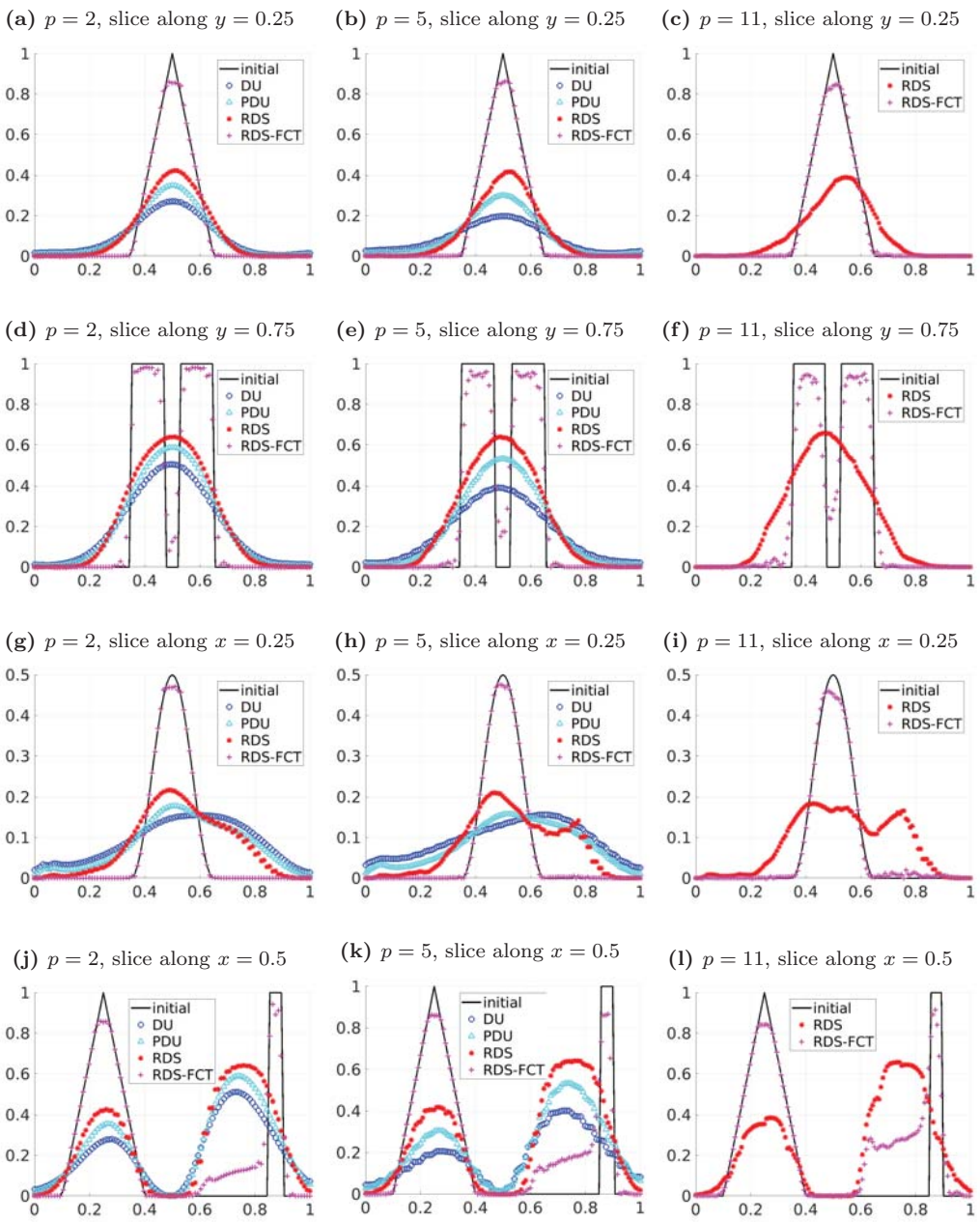

Fig. 6: Solid body rotation in 2D. Slices of the initial condition and numerical solutions $u_h(\cdot, 2\pi)$ obtained using low-order schemes and FCT for $p \in \{2, 5, 11\}$.

fact that low-order finite elements are better suited for numerical treatment of discontinuities than high-order approximations using the same #DOFs. The top of the smooth hump could be perfectly captured using the unlimited high-order DG scheme but the LED constraints of the FCT correction step do not distinguish between smooth and sharp peaks. Any change of the bound-preserving low-order solution at a local maximum or minimum is prevented by the limiter. In fact, even nonlinear schemes that strictly enforce preservation of local bounds can be at most second-order accurate in proximity to local extrema [39]. This order barrier can be circumvented using smoothness indicators like those proposed in [19, 30, 34].

8.3. Advection of a Gaussian hill in 2D

The matrix-based PDU scheme was found to produce p-independent results for the 1D examples of Section 8.1 and CG discretizations of the solid body rotation problem (as demonstrated by numerical studies in [34]). However, the DG results presented in Section 8.2 reveal that not only the DU scheme but also the PDU version may become more diffusive as p increases. To investigate possible reasons for this state of affairs, we perform additional studies of low-order schemes for the Gaussian initial condition

$$u_0(x,y) = \exp(-160((x-0.25)^2 + (y-0.25)^2). \tag{83}$$

In the first experiment, we translate u_0 with constant velocity $\mathbf{v}(x,y) = (1,1)$ up to the final time $T = 0.5$. In the second test, we use the velocity field of the solid body rotation benchmark and terminate computations at the time $T = \pi$. In both cases, the exact solution is given by

$$u(x,y,T) = \exp(-160((x-0.75)^2 + (y-0.75)^2). \tag{84}$$

The $L^1(\Omega)$ and $L^\infty(\Omega)$ errors for an increasing sequence of polynomial degrees p using (roughly) the same #DOFs are presented in Tables 1–2. These results confirm that the DU and RD schemes become more diffusive as p increases. The accuracy of PDU and RDS is virtually independent of p in the test in which the velocity is constant. In the solid body rotation test, p-refinements have a negative impact on the quality of PDU results. The only scheme which converges independently of p in both tests is RDS.

The unsatisfactory convergence behavior of DU for large p can be explained by the fact that each pair of nodes $i \in \mathcal{N}^e$ and $j \in \mathcal{N}^e \backslash \{i\}$ can

produce a diffusive flux of the form $d_{ij}^e(u_j^e - u_i^e)$, where d_{ij}^e is the artificial diffusion defined by (34). If i and j are arbitrary nodes of a high-order Bernstein finite element, the difference $u_j^e - u_i^e$ is likely to be greater than the difference between the DOFs of nearest neighbors. The only way to avoid large diffusive fluxes between nodes that lie far apart is to reduce the magnitude of $d_{ij}^e = \max\{-c_{ij}^e, 0, -c_{ji}^e\}$. As mentioned in Section 4, the PDU scheme achieves this by using the lumped-mass approximation P^eC^e to the Galerkin element matrix C^e. For constant velocity fields and CG discretizations on simplex meshes, the "preconditioner" $P^e = M_L^e(M_C^e)^{-1}$ converts C^e into an element matrix which has the same compact sparsity pattern as the one of the piecewise-linear subcell approximation with the same nodes [34]. Since the boundary element matrix S^e is treated separately, the modified element matrices $\tilde{C}^e = P^eC^e + \tilde{D}^e$ of the CG-PDU and DG-PDU schemes are the same (the difference lies in the way in which they are inserted into the global system). Hence, the analysis performed in [34] carries over to the DG version. It turns out that the suboptimal convergence behavior of PDU in the solid body rotation test is caused by the use of square elements. Contrary to the case of a simplicial mesh, the application of P^e does **not** make the element matrix C^e of the high-order Bernstein finite element discretization (approximately) sparse. However, the magnitude of off-diagonal entries associated with pairs of distant neighbors decreases, which explains the superiority of PDU over DU for large p. In Fig. 7, we present examples of element matrices C^e and P^eC^e for the two tests considered in this section.

The conclusion of this study is that none of the matrix-based schemes considered in this article is optimal in the sense that its convergence behavior is independent of p for general meshes and velocity fields. This is another reason that makes our new matrix-free RDS approach the method of choice when it comes to calculating low-order predictors for FCT algorithms.

8.4. Twisting rotation in 2D

In the final 2D example, we solve (1) in the domain $\Omega = (-1, 1)^2$ using $\mathbf{v}(x, y) = \frac{\pi}{2}d(x, y)^2(y, -x)^T$, where $d(x, y) = \max\{0, 1 - x^2\}\max\{0, 1 - y^2\}$. No boundary condition needs to be prescribed since d vanishes on Γ. In contrast to the solid body rotation benchmark, the velocity field is not divergence-free. It defines a twisting deformation of the initial condition

$$u_0(x, y) = \frac{1}{2}(\sin(\pi x)\sin(\pi y) + 1) \tag{85}$$

which is composed of four humps and attains values in the range [0,1].

		DU		PDU		RD		RDS	
p	#DOFs	L^1	L^∞	L^1	L^∞	L^1	L^∞	L^1	L^∞
1	$(2 \cdot 60)^2 = 14400$	1.27e−2	5.86e−1	1.18e−2	5.58e−1	9.78e−3	5.14e−1	9.78e−3	5.14e−1
2	$(3 \cdot 40)^2 = 14400$	1.48e−2	6.42e−1	1.18e−2	5.54e−1	1.50e−2	6.52e−1	9.23e−3	4.90e−1
3	$(4 \cdot 30)^2 = 14400$	1.62e−2	6.70e−1	1.17e−2	5.46e−1	1.88e−2	7.32e−1	9.01e−3	4.75e−1
4	$(5 \cdot 24)^2 = 14400$	1.73e−2	7.10e−1	1.17e−2	5.59e−1	2.15e−2	7.96e−1	9.05e−3	5.06e−1
5	$(6 \cdot 20)^2 = 14400$	1.82e−2	7.36e−1	1.17e−2	5.64e−1	2.35e−2	8.35e−1	9.05e−3	5.17e−1
6	$(7 \cdot 17)^2 = 14161$	1.90e−2	7.19e−1	1.17e−2	5.34e−1	2.51e−2	8.25e−1	9.26e−3	4.89e−1
7	$(8 \cdot 15)^2 = 14400$	1.96e−2	7.44e−1	1.16e−2	5.42e−1	2.63e−2	8.59e−1	9.33e−3	4.96e−1
8	$(9 \cdot 13)^2 = 13689$	2.03e−2	7.51e−1	1.17e−2	5.43e−1	2.75e−2	8.62e−1	9.49e−3	4.97e−1
9	$(10 \cdot 12)^2 = 14400$	2.05e−2	7.89e−1	1.15e−2	5.68e−1	2.83e−2	9.04e−1	9.52e−3	5.50e−1
10	$(11 \cdot 11)^2 = 14641$	2.10e−2	7.55e−1	1.31e−2	5.74e−1	2.90e−2	8.74e−1	9.66e−3	4.89e−1

Table 1: Translation of a Gaussian hill, comparative study of low-order methods for $p = 1, \ldots, 10$.

		DU		PDU		RD		RDS	
p	#DOFs	L^1	L^∞	L^1	L^∞	L^1	L^∞	L^1	L^∞
1	$(2 \cdot 60)^2 = 14400$	1.55e−2	6.63e−1	1.45e−2	6.35e−1	1.30e−2	6.08e−1	1.30e−2	6.08e−1
2	$(3 \cdot 40)^2 = 14400$	1.79e−2	7.14e−1	1.47e−2	6.37e−1	1.81e−2	7.20e−1	1.18e−2	5.64e−1
3	$(4 \cdot 30)^2 = 14400$	1.94e−2	7.38e−1	1.51e−2	6.38e−1	2.20e−2	7.87e−1	1.16e−2	5.54e−1
4	$(5 \cdot 24)^2 = 14400$	2.06e−2	7.75e−1	1.56e−2	6.68e−1	2.46e−2	8.43e−1	1.18e−2	5.83e−1
5	$(6 \cdot 20)^2 = 14400$	2.16e−2	7.99e−1	1.63e−2	6.90e−1	2.66e−2	8.77e−1	1.19e−2	5.98e−1
6	$(7 \cdot 17)^2 = 14161$	2.24e−2	7.80e−1	1.72e−2	6.78e−1	2.80e−2	8.62e−1	1.22e−2	5.72e−1
7	$(8 \cdot 15)^2 = 14400$	2.30e−2	8.01e−1	1.79e−2	7.04e−1	2.92e−2	8.89e−1	1.23e−2	5.82e−1
8	$(9 \cdot 13)^2 = 13689$	2.38e−2	8.07e−1	1.90e−2	7.23e−1	3.02e−2	8.90e−1	1.27e−2	5.96e−1
9	$(10 \cdot 12)^2 = 14400$	2.41e−2	8.46e−1	1.95e−2	7.66e−1	3.08e−2	9.31e−1	1.26e−2	6.38e−1
10	$(11 \cdot 11)^2 = 14641$	2.45e−2	8.07e−1	2.13e−2	7.56e−1	3.13e−2	8.96e−1	1.29e−2	5.90e−1

Table 2: Solid body rotation of a Gaussian hill, comparative study of low-order methods for $p = 1, \ldots, 10$.

$$10^{-2} \cdot \begin{bmatrix} 1.67 & -0.14 & -0.14 & -0.14 & -0.56 & -0.23 & -0.14 & -0.23 & -0.09 \\ 0.97 & 0.56 & -0.14 & -0.00 & -0.37 & -0.56 & -0.05 & -0.19 & -0.23 \\ 0.42 & 0.97 & 0.00 & 0.05 & -0.00 & -0.97 & -0.00 & -0.05 & -0.42 \\ 0.97 & -0.00 & -0.05 & 0.56 & -0.37 & -0.19 & -0.14 & -0.56 & -0.23 \\ 0.56 & 0.37 & 0.00 & 0.37 & -0.00 & -0.37 & -0.00 & -0.37 & -0.56 \\ 0.23 & 0.56 & 0.14 & 0.19 & 0.37 & -0.56 & 0.05 & -0.00 & -0.97 \\ 0.42 & 0.05 & 0.00 & 0.97 & -0.00 & -0.05 & 0.00 & -0.97 & -0.42 \\ 0.23 & 0.19 & 0.05 & 0.56 & 0.37 & -0.00 & 0.14 & -0.56 & -0.97 \\ 0.09 & 0.23 & 0.14 & 0.23 & 0.56 & 0.14 & 0.14 & 0.14 & -1.67 \end{bmatrix}$$

(a) DU matrix C^e for velocity field $\mathbf{v} = (1, 1)^T$.

$$10^{-2} \cdot \begin{bmatrix} 3.70 & -1.85 & 0.00 & -1.85 & -0.00 & 0.00 & 0.00 & 0.00 & -0.00 \\ 0.93 & 1.85 & -0.93 & -0.00 & -1.85 & -0.00 & 0.00 & -0.00 & 0.00 \\ -0.00 & 1.85 & 0.00 & 0.00 & -0.00 & -1.85 & -0.00 & 0.00 & -0.00 \\ 0.93 & -0.00 & 0.00 & 1.85 & -1.85 & -0.00 & -0.93 & -0.00 & 0.00 \\ -0.00 & 0.93 & 0.00 & 0.93 & -0.00 & -0.93 & 0.00 & -0.93 & 0.00 \\ 0.00 & 0.00 & 0.93 & -0.00 & 1.85 & -1.85 & 0.00 & -0.00 & -0.93 \\ -0.00 & -0.00 & -0.00 & 1.85 & -0.00 & 0.00 & 0.00 & -1.85 & -0.00 \\ 0.00 & 0.00 & -0.00 & -0.00 & 1.85 & -0.00 & 0.93 & -1.85 & -0.93 \\ -0.00 & -0.00 & -0.00 & 0.00 & -0.00 & 1.85 & 0.00 & 1.85 & -3.70 \end{bmatrix}$$

(b) PDU matrix $P^e C^e$ for velocity field $\mathbf{v} = (1, 1)^T$.

$$10^{-2} \cdot \begin{bmatrix} -0.46 & 0.76 & 0.01 & -0.30 & -0.00 & -0.01 & -0.01 & 0.01 & 0.00 \\ -0.35 & 0.39 & 0.35 & -0.03 & -0.39 & 0.03 & -0.00 & 0.00 & 0.00 \\ -0.01 & -0.76 & 1.23 & 0.01 & 0.00 & -0.47 & -0.00 & -0.01 & 0.01 \\ 0.12 & 0.03 & 0.00 & -0.85 & 0.85 & -0.00 & -0.12 & -0.03 & -0.00 \\ -0.00 & 0.19 & -0.00 & -0.42 & -0.00 & 0.42 & -0.00 & -0.19 & 0.00 \\ -0.00 & -0.03 & 0.27 & -0.00 & -0.85 & 0.85 & 0.00 & 0.03 & -0.27 \\ 0.01 & -0.01 & -0.00 & 0.30 & -0.00 & 0.01 & -1.23 & 0.93 & -0.01 \\ 0.00 & -0.00 & -0.00 & 0.03 & 0.39 & -0.03 & -0.50 & -0.39 & 0.50 \\ -0.00 & 0.01 & -0.01 & -0.01 & 0.00 & 0.47 & 0.01 & -0.93 & 0.46 \end{bmatrix}$$

(c) PDU matrix $P^e C^e$ for velocity field $\mathbf{v}(x, y) = (0.5 - y,\ x - 0.5)^T$.

Fig. 7: Advection matrices of a biquadratic Bernstein element ($p = 2$, $h = \frac{1}{12}$, lexicographical node numbering). Entries in gray vanish in the sparse matrix assembled from subcell contributions of the piecewise-bilinear Bézier net approximation.

At the final time $T = 9$, the swirling flow produces a snail-like deformed

configuration consisting of four intertwined spirals. The exact solution is constant along the characteristics of the linear advection equation and, therefore, attains the same maxima and minima as the initial data. Due to the narrow gaps between adjacent spirals and a complex non-solenoidal velocity field, this example is well suited for testing the ability of numerical schemes to resolve fine-scale features in a crisp and nonoscillatory manner.

In Fig. 9, we show the results obtained with DU, PDU, RDS, and RDS-FCT. Simulations are performed with $p \in \{1, 3\}$ for all schemes and, additionally, $p \in \{7, 15\}$ for matrix-free schemes. The time step is set to $\Delta t = 2.5 \cdot 10^{-3}$. All methods under investigation produce bound-preserving numerical approximations, and their relative performance is similar to that observed in previous examples. The accuracy of DU deteriorates as p increases, while the convergence behavior of PDU and RDS is largely independent of p. In terms of accuracy, RDS outperforms PDU by a wider margin in this test. In contrast to the strongly smeared PDU results, the RDS solutions are relatively well resolved, and their global maxima/minima are closer to those of the exact solution. The distortions in the shapes of the RDS(7) and RDS(15) solutions are caused by coarse mesh resolution and become less pronounced on finer meshes. We intentionally restrict ourselves to simulations with relatively few global DOFs because the objective of our study is to identify schemes that yield the best coarse-level approximations.

Once again, RDS turns out to be the only low-order method which produces satisfactory results independently of p. Moreover, the possibility of a matrix-free implementation and the simple formula for the distribution weights makes it very efficient. The solutions produced by the RDS-FCT version are remarkably crisp and free of numerical artifacts that were observed in the RDS results for $p \in \{7, 15\}$. Moreover, the high quality of the FCT solutions is not significantly affected by p-refinements. Further improvements of the limiting strategy (localization to subcells, use of smoothness indicators) may lead to FCT schemes that recover the underlying high-order approximation exactly for problems with sufficiently smooth solutions.

8.5. Advection of "balls and jacks" in 3D

The last test problem that we consider in this article is the 3D advection of "balls and jacks" with constant velocity $\mathbf{v}(x, y, z) = (5, 5, 5)$ in $\Omega = (0, 100)^3$ and homogeneous boundary conditions at the inlet. The simulation ends at the final time $T = 8$. The piecewise-constant initial condition [8]

$$u_0 = u_1 + 2u_2 + 3u_3,$$

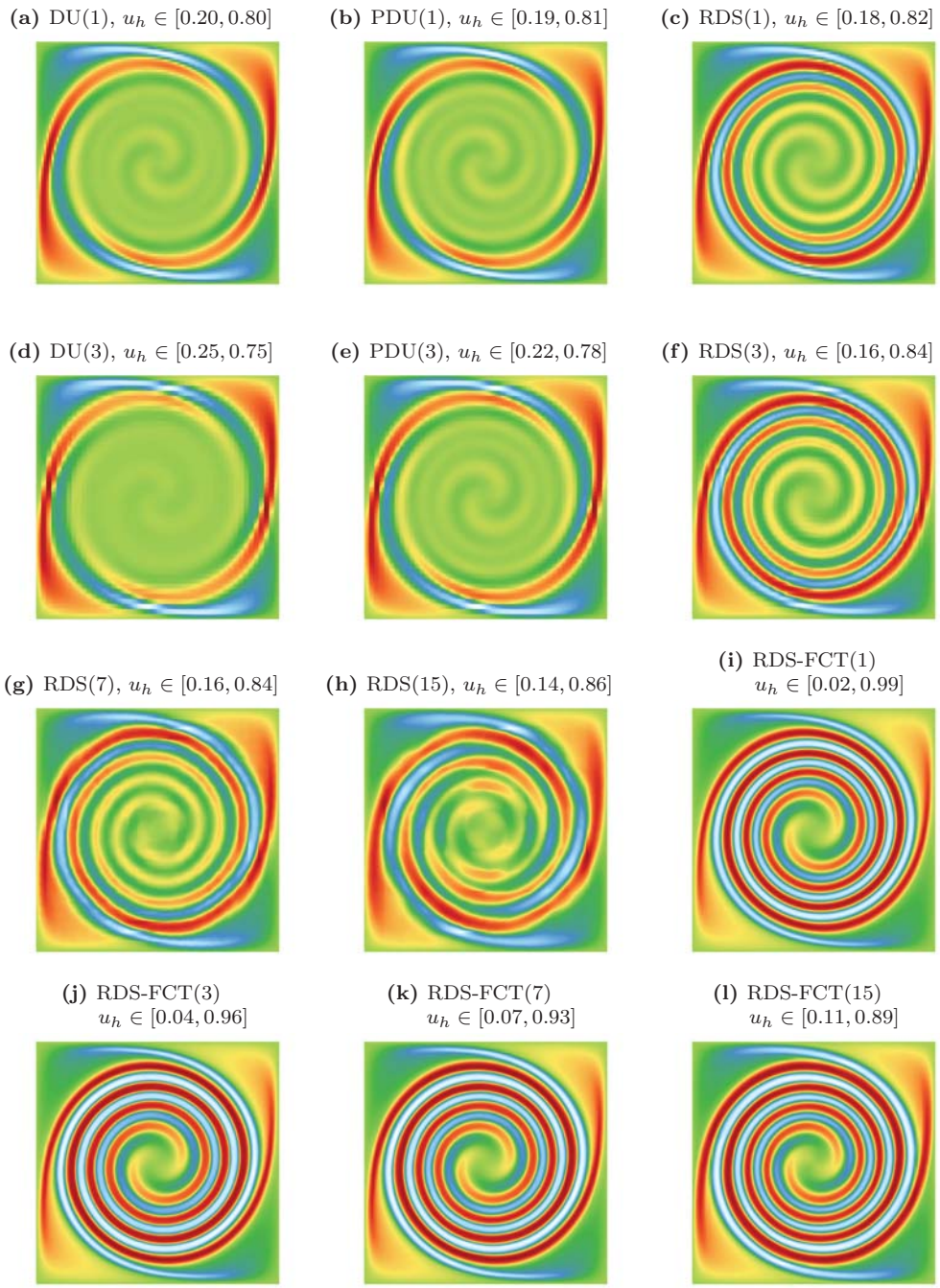

Fig. 9: Twisting rotation in 2D. Numerical solutions $u_h(\cdot, 9)$ for $p \in \{1, 3, 7, 15\}$.

as shown in Fig. 10a–10c, is defined using the characteristic functions

$$u_i(x,y,z) = \begin{cases} 1 & \text{if } (x,y,z) \in \Omega_i, \\ 0 & \text{otherwise,} \end{cases}$$

of the subdomains Ω_i, $i = 1,2,3$ with $\Omega_1 = \Omega \backslash (\Omega_2 \cup \Omega_3)$. The subdomain Ω_2 is composed of the 3D cross $\{(x,y,z) \in (7,32) \times (10,13) \times (10,13) \cup (14,17) \times (3,26) \times (10,13) \cup (14,17) \times (10,13) \times (3,26)\}$ rotated by -45 degrees in the xy-plane, the shell (i.e., the difference of two balls) centered at $(x,y,z) = (40,20,20)$ with radii 3 and 7, and the shell centered at $(x,y,z) = (40,40,40)$ with radii 7 and 10. The subdomain Ω_3 combines the 3D cross $\{(x,y,z) \in (2,27) \times (30,33) \times (30,33) \cup (9,12) \times (23,46) \times (30,33) \cup (9,12) \times (30,33) \times (23,46)\}$, the ball centered at $(x,y,z) = (40,20,20)$ with radius 3, the ball centered at $(x,y,z) = (40,40,40)$ with radius 7, and the shell centered at $(x,y,z) = (40,20,20)$ with radii 7 and 10.

Similarly to the example of Section 8.2, the challenge of this test is to preserve the shape of u_0 in the process of advection from one corner of the box Ω into another. Numerical solutions are calculated using $N_h^{\mathrm{DG}} = 192^3$ and the time step $\Delta t = 10^{-3}$. The RDS and RDS-FCT results for $p \in \{1,3,7\}$ are shown in Fig. 10. For comparison purposes, we also present the DU(1) and PDU(1) results in Fig. 11. While all low-order solutions are significantly diffused, the balls and jacks are still recognizable. The range of u_h is displayed above each plot to quantify and compare the levels of numerical diffusion. Even for $p = 1$, the PDU scheme is more accurate than DU but inferior to RDS. Since the velocity is constant in this experiment, the PDU matrices $\tilde{C}^e$ for $p > 1$ have the same sparsity pattern as the element matrix of the corresponding piecewise-trilinear subcell approximation. Therefore, the PDU scheme would be a viable alternative to RDS in this test. However, the matrix-free nature and p-independent convergence behavior of the RDS scheme make it a better choice for general 3D advection problems.

The combination of RDS with FCT produces much better results for the 3D advection of "balls and jacks" than the DU-FCT algorithm developed in [8]. Remarkably, the RDS-FCT solutions presented in Fig. 10 are free of numerical artifacts that motivated the use of restrictive local bounds in [8]. In our experience, the use of extended bounds is admissible and has a positive impact on the accuracy of FCT solutions if these bounds are defined using data from a properly chosen element-stencil neighborhood of a node. In contrast to the extended bounds that were found to perform so poorly

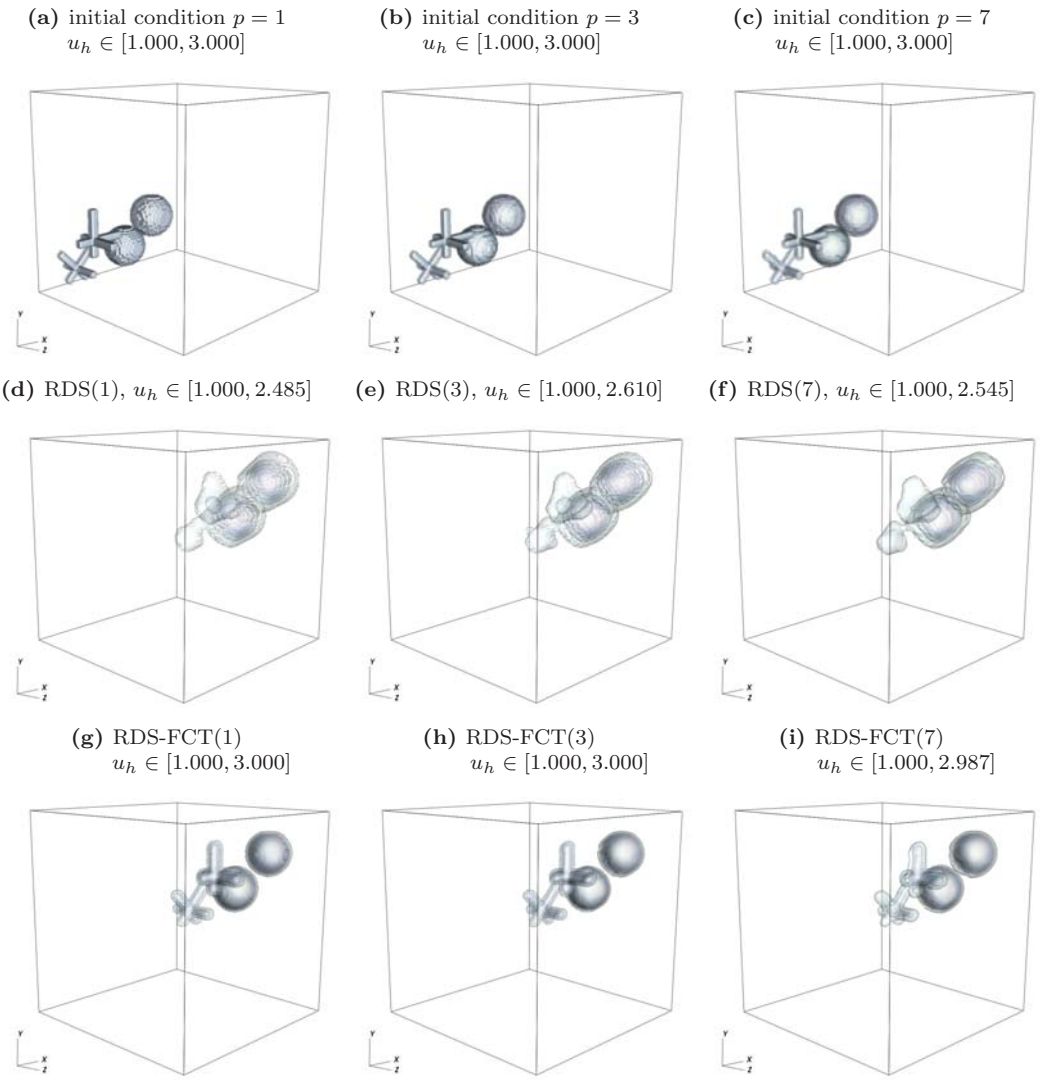

Fig. 10: Advection of "balls and jacks" in 3D. Initial condition and numerical (RDS, RDS-FCT) solutions $u_h(\cdot, 8)$ for $p \in \{1, 3, 7\}$, volume rendering with transparency.

in [8], the local extrema defined by (19) use data from elements meeting at the point $\mathbf{x}_i$ rather than the local DOFs of the element K^e to which u_i^e belongs and the data in common edge/face neighbors. The high quality of the DG results obtained with definition (19) of the local bounds for the FCT correction step confirms the findings of [34], where the use of element-stencil bounds was shown to be preferable in the context of CG-FCT algorithms.

(a) DU(1), $u_h \in [1.000, 2.125]$ **(b)** PDU(1), $u_h \in [1.000, 2.311]$ **(c)** RDS(1), $u_h \in [1.000, 2.485]$

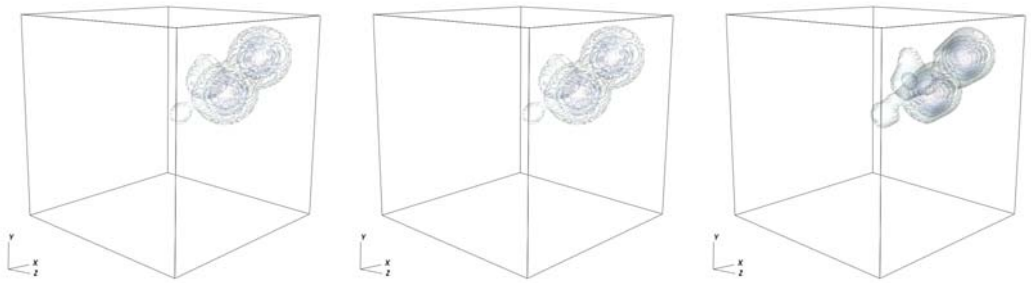

Fig. 11: Advection of "balls and jacks" in 3D. Low-order (DU, PDU, RDS) numerical solutions $u_h(\cdot, 8)$ for $p = 1$, volume rendering with transparency.

9. Conclusions

The residual distribution framework proposed in this work extends classical RD approaches to DG methods and localizes them to element subcells in a simple way. The resulting nonlinear low-order schemes are provably local extremum diminishing and can be implemented in a matrix-free manner. Their use in FCT algorithms for high-order Bernstein finite element approximations produces excellent results for the time-dependent advection problems considered in the presented numerical study. Preliminary investigations indicate that the same methodology can be used to construct bound-preserving finite element schemes for anisotropic diffusion equations and stationary problems. Work is under way to combine residual distribution ideas with monolithic limiting approaches based on the theory developed in [12, 33]. The outcomes of these efforts will be reported in a forthcoming publication.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy under Contract DE-AC52-07NA27344, LLNL-JRNL-768125. The first two authors were supported by the German Research Association (DFG) under grant KU 1530/23-1. The collaboration with all coauthors of [8, 34] at early stages of this research is gratefully acknowledged. A special thanks to Christoph Lohmann (TU Dortmund University) for careful proofreading of the paper and helpful suggestions. Last but not least, the authors would like to thank Svetlana Tokareva (Los Alamos National Laboratory) for discussions leading to a deeper understanding of relationships to existing residual distribution schemes.

References

[1] R. Abgrall, Essentially non-oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys.*, **214** (2006) 773–808.

[2] R. Abgrall, A review of residual distribution schemes for hyperbolic and parabolic problems: The July 2010 state of the art. *Commun. Comput. Phys.* **11** (2012) 1043–1080.

[3] R. Abgrall, P. Bacigaluppi and S. Tokareva, How to avoid mass matrix for linear hyperbolic problems. In: B. Karasözen et al. *Numerical Mathematics and Advanced Applications (ENUMATH 2015)*. Springer, 2016, 75–86.

[4] R. Abgrall and P.L. Roe, High order fluctuation schemes on triangular meshes. *J. Sci. Comput.* **19** (2003) 3–36.

[5] R. Abgrall and S. Tokareva, Staggered grid residual distribution scheme for Lagrangian hydrodynamics. *SIAM J. Sci. Comput.* **39** (2017) A2317–A2344.

[6] R. Abgrall and J. Trefilík, An example of high order residual distribution scheme using non-Lagrange elements. *J. Sci. Comput.* **45** (2010) 3–25.

[7] R. Abgrall, Q. Viville, H. Beaugendre and C. Dobrzynski, Construction of a p-adaptive continuous residual distribution scheme. *J. Sci. Comput.* **72** (2017) 1232–1268.

[8] R. Anderson, V. Dobrev, Tz. Kolev, D. Kuzmin, M. Quezada de Luna, R. Rieben and V. Tomov, High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation. *J. Comput. Phys.* **334** (2017) 102–124.

[9] R. Anderson, V. Dobrev, Tz. Kolev, R. Rieben, and V. Tomov, High-order multi-material ALE hydrodynamics, *SIAM J. Sci. Comput.* **40** (2018) B32–B58.

[10] S. Badia and J. Bonilla, Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Computer Methods Appl. Mech. Engrg.* **313** (2017) 133-158.

[11] S. Badia, J. Bonilla and A. Hierro, Differentiable monotonicity-preserving schemes for discontinuous Galerkin methods on arbitrary meshes. *Computer Methods Appl. Mech. Engrg.* **320** (2017 582-605.

[12] G. Barrenechea, V. John and P. Knobloch, Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.* **54** (2016) 2427–2451.

[13] Ch. Bernardi, Th. Chacón Rebollo and M. Restelli, A posteriori analysis of a positive streamwise invariant discretization of a convection-diffusion equation. *J. Sci. Comput.* **51** (2012) 349–374.

[14] J. Brown, J.-S. Camier, V. Dobrev, P. Fischer, Tz. Kolev, T. Ratnayaka, M. Shephard, J. Thompson, and V. Tomov, CEED-MS10 milestone report: Initial CEED API, *Center for efficient exascale discretization (CEED), Exascale computing project* **DOI:10.5281/zenodo.2542340** (2017).

[15] J.-C. Carette, H. Deconinck, H. Paillère and P.L. Roe, Multidimensional upwinding: Its relation to finite elements. *Int. J. Numer. Methods Fluids* **20**:8-9 (1995) 935–955.

[16] C.J. Cotter and D. Kuzmin, Embedded discontinuous Galerkin transport schemes with localised limiters. *J. Comput. Phys.* **311** (2016) 363–373.

[17] H. Deconinck, H. Paillère, R. Struijs and P.L. Roe, Multidimensional upwind schemes based on fluctuation-splitting for systems of conservation laws. *Comput. Mech.* **11**:5-6 (1993) 323–340.

[18] H. Deconinck and M. Ricchiuto, Residual distribution schemes: foundations and analysis. *Encyclopedia of Computational Mechanics*, Volume 3: Fluids, Wiley, 2007.

[19] S. Diot, S. Clain and R. Loubére, Improved detection criteria for the Multi-dimensional Optimal Order Detection (MOOD) on unstructured meshes with very high-order polynomials. *Comput. Fluids* 64 (2012) 43–63.

[20] D. A. Di Pietro and A. Ern, Mathematical Aspects of Discontinuous Galerkin Methods. *Math & Applications* **69**, Springer-Verlag, Berlin, 2012.

[21] M. Dumbser, O. Zanotti, R. Loubère and S. Diot, A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *J. Comput. Phys.* **278** (2014) 47–75.

[22] S.K. Godunov, Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.* **47** (1959) 271–306.

[23] S. Gottlieb, C.-W. Shu and E. Tadmor, Strong stability-preserving high-order time discretization methods. *SIAM Review* **43** (2001) 89–112.

[24] T. N. T. Goodman, Further variation diminishing properties of Bernstein polynomials on triangles. *Constructive Approximation* **3** (1987) 297–305.

[25] J.-L. Guermond, M. Nazarov, B. Popov and Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. *SIAM Journal on Numerical Analysis* **52** (2014) 2163–2182.

[26] R.C. Kirby, Fast inversion of the simplicial Bernstein mass matrix. *Numer. Math.* **135** (2017) 73–95.

[27] L. Krivodonova, Limiters for high-order discontinuous Galerkin methods. *J. Comput. Phys.* **226** (2007) 879–896.

[28] D. Kuzmin, A vertex-based hierarchical slope limiter for p-adaptive discontinuous Galerkin methods. *J. Comput. Appl. Math.* **233** (2010) 3077–3085.

[29] D. Kuzmin, Algebraic flux correction I. Scalar conservation laws. In: D. Kuzmin, R. Löhner, S. Turek (eds), *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Springer, 2nd edition, 2012, pp. 145–192.

[30] D. Kuzmin, M. Quezada de Luna and C. Kees, A partition of unity approach to adaptivity and limiting in continuous finite element methods. *Ergebnisber. Angew. Math.* **590**, TU Dortmund University, 2018.

[31] D. Kuzmin and S. Turek, Flux correction tools for finite elements. *J. Comput. Phys.* **175** (2002) 525–558.

[32] R.J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis* **33**, (1996) 627–665.

[33] C. Lohmann, *Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems.* PhD thesis, TU Dortmund University, 2019.

[34] C. Lohmann, D. Kuzmin, J.N. Shadid and S. Mabuza, Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. *J. Comput. Phys.* **344** (2017) 151-186.

[35] MFEM: Modular finite element methods. https://mfem.org.

[36] R. Struijs, Multi-dimensional upwind discretization method for the Euler equations on unstructured grids. *PhD thesis* (1994)

[37] F. Vilar, A posteriori correction of high-order discontinuous Galerkin scheme through subcell finite volume formulation and flux reconstruction. *J. Comput. Phys.* **387** (2019) 245–279.

[38] X. Zhang and C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.* **229** (2010) 3091–3120.

[39] X. Zhang and C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. A* **467** (2011) 2752–2776.